PERSPECTA

Volume 5, Number 1, 2015
5. kötet, 1. szám, 2015

RESULTS OF AN ONLINE COMPLEX VOCABULARY TEST ASSESSING YOUNG LEARNERS' ENGLISH AS A FOREIGN LANGUAGE WORD KNOWLEDGE

Istvan Thekes

Abstract. The learning and teaching of vocabulary is a popular research area in the FL learning literature which is in the center of attention of both scholars and teachers. Educators have been encouraged (Lewis, 1993; Thornbury, 2002) to promote intentional learning of words in the classroom. Since the early 1990s textbook and FL syllabus writers have laid special emphasis on building vocabulary in the curriculum (Fitzpatrick, Al-Qarni & Meara, 2008). Knowing a word is complex and multidimensional in nature. Various aspects of knowing a word must be considered. Breadth of vocabulary knowledge means how many words a person knows while depth refers to the knowledge of dimensions, e.g. synonyms, antonyms, contextual use, etc. A diagnostic complex vocabulary test was designed to assess learners' word knowledge. Most of the diagnostic vocabulary tests measure one dimension of vocabulary (Nation, 1990). They either tap into receptive or productive word knowledge; thus there is a need to develop a diagnostic online English as a FL vocabulary test assessing Young learners' receptive and productive word knowledge. Findings indicate that word knowledge is a complex construct. Students overall performed better on the receptive vocabulary tasks. Results are analyzed in depth with regard to the triangulated data. Classroom implications and limitations are also discussed.

Keywords: foreign language vocabulary, assessment, young learner

1. Introduction

The 1990s saw an increasing number of studies focusing on foreign language (FL) vocabulary learning and the literature has been growing ever since then extending the knowledge on such areas as FL vocabulary assessment (Nation, 2001; Laufer, Elder, Congdon & Hill, 2004; Schmitt, 1998), the FL mental lexicon (Singleton, 1999; Zareva, 2007), corpus studies (Horváth, 2001; Kilgarriff, 1997; Nation & Macalister, 2010) and vocabulary learning strategies (Chostelidou, Griva, Ioannidis & Panitsidou, 2012; Nation, 1990; Schmitt, 2000). Doró (2013, p. 5) claims that emerging questions have only been partly elaborated on and new questions arise in an increasing number. It has also been affirmed that vocabulary knowledge is a good predictor of reading comprehension (Nassaji, 2004; Shiotsu & Weir, 2007) and general language proficiency (Zimmerman, 2004). It has been reported that receptive vocabulary knowledge predicts productive vocabulary knowledge (Laufer & Nation, 1999, p. 42). With the emergence of the lexical approach (Lewis, 1993) in language teaching, vocabulary research gained even more importance. A practitioner uncompromisingly concludes "Without grammar very little can be conveyed, without vocabulary nothing can be conveyed" (Thornbury, 2002, p. 18).

The learning and teaching of vocabulary is a popular research area in the FL learning literature which is in the center of attention of both scholars and teachers. Educators have been encouraged (Lewis, 1993; Thornbury, 2002) to promote intentional learning of words in the classroom. Since the early 1990s textbook and FL syllabus writers have laid special emphasis on building vocabulary in the curriculum (Fitzpatrick, Al-Qarni & Meara, 2008). Successful language learning is greatly determined by FL word knowledge (Schoonen & Verhallen, 2008). The field of vocabulary and word knowledge is researched by several actors in scientific domains. Education researchers (Nagy, 2004), psycholinguists (Ellis & Beaton, 1997), neurolinguists (Paradis, 2004), morphologists (Jackson & Zé

Received 15 September 2015.

Amvela, 2011) and language teachers (Thékes, 2014; Thornbury, 2002) contribute to or exploit the empirical results of vocabulary learning research and assessment.

Vocabulary is nowadays considered as one of the strongest predictors of FL proficiency (Schmitt, 2008, p. 352). Diagnostic vocabulary tests have been developed and validated in the past 30 years. The major characteristic feature of these instruments is that they test one dimension of knowledge. They either assess receptive or productive knowledge of words and hardly any attempt has been made towards an instrument that assesses both receptive and productive FL word knowledge.

2. Theoretical background

2.1. Aspects of knowing a word in a foreign language

Knowing a word is complex and multidimensional in nature. Various aspects of knowing a word must be considered. Breadth of vocabulary knowledge means how many words a person knows while depth refers to the knowledge of dimensions, e.g. synonyms, antonyms, contextual use, etc. The complexity of the concept of knowing a word is emphasized by Mukarto (2005, p. 153) who declares that

„...learning even one FL word or a lexical item is a complex task. Naturally, learners' knowledge of a word is not binary in nature, nor is it an all or nothing phenomenon."

Several dimensions have been identified that inform researchers and teachers how complex it is to determine what it means to know a word. When considering Young learners, it must be kept in mind that word knowledge is incremental, which implies multiple oral and written inputs (Nagy, Anderson & Herman, 1987, p. 238). Word knowledge is also multidimensional since a lot of words have different meanings. Finally, word knowledge is interrelated in that the knowledge of one lexical item is connected to another (Scott & de la Fuente 2008, p. 108).

Bogaards (2000, p. 146) further claims that FL learners may learn the subsequent dimensions: form (spoken and written), meaning, morphology, syntax, collocates and discourse. Nagy and Scott (2000, p. 278) identified several dimensions that describe the complexity of what it means to know a word. First, word knowledge is incremental, which involves many encounters with both spoken and written words in varying contexts (Nagy et al., 1987). Second, word knowledge is multidimensional because many words have multiple meanings and serve different functions in different contexts. Third, word knowledge is interrelated in that knowledge of one word connects to knowledge of other words.

2.2. Diagnostic assessment in the context of word knowledge of young learners

It is postulated by Read (2000, p. 32) that there are two contrasting perspectives of vocabulary assessment. One viewpoint is that vocabulary items can be tested as an independent semantic field independent of context. The other view is that lexis must always be measured in context.

This issue should be the concern of test-givers. The issues emerging from language testing research need to be looked at from four different angles (Lehmann, 2009; Nation, 2013; Milton & Fitzpatrick, 2014). Four major questions are proposed by Nation (2013) that need to be addressed: (1) why to test vocabulary? (2) what words to test? (3) what aspects of word knowledge to test? (4) how to test the various aspects of word knowledge?

Before presenting the data-gathering instrument, and the findings of studies assessing the word knowledge of Young learners, I will elaborate on the characteristic traits and principles of diagnostic testing of FL in the context of Young learners. Nikolov and Szabó (2011). These principles had been grounded the study by Alderson (2005) and McKay (2006, p. 36) who states that diagnostic assessment of Young learners' FL proficiency and word knowledge is important because a lot can be implied from them in classroom practices.

I will attempt to synthesize six of these traits which, I believe, are the most relevant from the perspective of computerized vocabulary assessment of young learners: (1) the purpose of diagnostic tests is to identify the strengths and weaknesses of learners; (2) diagnostic tests must make it possible to analyze the score of each item in detail and to report the results; thus they provide feedback in detail and further steps can be taken; (3) diagnostic tests are low-stakes tests or consequences are of irrelevant weight so optimal achievement is not hindered by anxiety or any other affective factor; (4)

diagnostic tests are more likely to focus on 'lower-level' linguistic abilities than on 'higher-level' abilities; (5) diagnostic testing is probably made more efficient by using a computerized platform.

2.3. Foreign language vocabulary tests

Ever since vocabulary became the point of convergence of foreign language learning studies, assessment of word knowledge has been regarded as a fundamental issue in the research of this domain. Special attention will be drawn to (1) the computerized versions of these instruments as in the 21st century diagnostic assessment is predominantly executed in an online environment (Laufer et al., 2004) and (2) whether the data collecting instruments to be discussed have versions designed for Young learners. It must also be highlighted that there is a consensus among scholars in foreign language vocabulary assessment that various modalities of item assessment exist. Laufer et al. (2004, p. 218) claim that words may be measured from two perspectives: (1) form-focused perspective that implies that the test-taker is able to retrieve the form of the word evidencing productive knowledge; (2) meaning-focused perspective that entails that the test-taker can retrieve the form of the word evidencing receptive knowledge. Laufer et al. (2004) refer to the productive-receptive dichotomy as active-passive knowledge. The scholars distinguish among four degrees of knowledge of meaning, on the basis of two dichotomous distinctions: providing the form for a given concept vs. providing the meaning for a given form; and recall vs. recognition (of form or meaning). These distinctions entail the following four modalities constituting a hierarchy of difficulty: (1) passive recognition that involves recognizing an item in a multiple choice test; (2) active recognition that includes a given definition and four items; in this modality the definition must be matched with the pertaining item; (3) passive recall that incorporates a sentence and the synonym of one item in the sentence must be given by the test-takers; and (4) active recall that comprises a description of an item and the initial letter of the item is provided; test-takers are expected to produce the word. In a review article Schmitt (2014, p. 921) uses different terms for the same concepts. Passive recognition is termed meaning recognition; active recognition is named form recognition whereas passive recall is termed meaning recall and active recall is called form recall. In an attempt to provide examples, sample tasks will be given subsequently.

2.4. Vocabulary Levels Test

The Vocabulary Levels Test (VLT) is a receptive vocabulary test with a discrete point measure. It requires meaning recognition. The test was developed by Nation (1990) and it was validated by Schmitt, Schmitt and Clapham (2001). Words are selected from such corpora as British National Corpus (Kilgarriff, 1997) and Cambridge and Nottingham Corpus of Discourse in English (CANCODE) up to five levels: the first 2,000, 3,000, 5,000 and 10,000 most frequent words. These levels bear importance from a research-based perspective. The 2,000-3,000 levels contain high-frequency words whose knowledge is necessary for everyday communication. The 5,000 level is the minimal size which learners can understand authentic texts with. The 10,000 level, contains the most common low-frequency words (Webb & Sassao 2013). The fifth level is not grounded on any corpus but includes items from the University Word List (Xue & Nation, 1984). The test-taker sees six words on the left-hand side and three definitions or synonyms on the right-hand side. They are expected to match the right-hand side items with three of the sox words on the left-hand side. This means that the task contains three distractors. In the entire test each level comprises six clusters of six words. Table 1 presents one sample task of the VLT.

Table 1. Sample task of the VLT

Instruction: match three of the words from 1) to 6) with three definitions A) - C)

1 bitter		
2 independent	A)	very small
3 lovely	B)	beautiful
4 merry	C)	liked by many people
5 popular		
6 slight		

Since the test gives estimates of vocabulary size at 5 levels, it can be applied for placement purposes and for diagnosis of vocabulary gaps. Four parallel test versions were developed. The criterion of the development of the test was that the definitions are short; the test could be done in the fastest possible time and with the proper arrangement of the possibility of blind guesses could be diminished. In the online version of the VLT the test-taker is expected to write the listed six words next to the three definitions. The evaluation of the test is automatically done after the test. With the modified version of the online test, Vocabulary Online Recognition Speed Test (VORST) the speed of word recognition can also be examined (Laufer & Nation, 2001, p.21).

A version of the VLT designed for Young learners has also been developed. Catalan Jimenez and Terrazas Gallego (2008) used this version of the instrument with young Spanish Young learners of English. They modified the word selection process by involving such low-frequency words as names of animals (e.g., 'lion', 'ostrich', 'tiger') that Young learners might know better than high-frequency words used by adults (e.g., 'beer', 'office', 'wine'). The researchers reported that the Young learners' version of the VLT proved to be a valid measure of vocabulary assessment.

2.5. Productive Vocabulary Levels Test

With regard to productive knowledge of vocabulary, Laufer and Nation (1995) developed an instrument that measures productive word knowledge. The test requires form recall from the part of the participants. Similarly to the Vocabulary Levels Test, the tasks are divided into frequency clusters: 2,000, 3,000, 5,000, 10,000. In this test students see sentences. In each sentence only the initial letters of one word are given. Students must write the missing part of the word. This test is originally named the Test of Controlled Productive Ability (TCPA), nowadays it is referred to as Productive Vocabulary Levels Test (PVLT). A part of the instrument is presented in Table 2.

Table 2. Productive Vocabulary Levels Test

Instruction: Complete the words by filling in the gaps with the proper letters
He likes walking in the fo…… because the trees are beautiful there.
1) He takes cr.....and sugar in his coffee
The actor took the st…… to perform in the long-awaited play.

It is obvious from Table 3, that the sentences following one another are unrelated. The test format resembles a C-test to a great extent. In the pilot study of the instrument the researchers reported that the selection of the target words was determined with the purpose of avoiding any ambiguity of the meaning of the word. Similar to Schmitt et al. (2001) four test versions were developed. It must be noted that the test has been criticized from a construct validity point of view. It was pointed out by Read (2000, p. 66) that the instrument might not assess productive word knowledge. He argues that some of the items demand only recognition and some of them need more contextual clues than others, thus he is dubious whether the test assesses what it is meant to assess.

Abduallah, Puteh, Azizan, Hamdan & Saude (2013) used the PVLT to assess the productive vocabulary of 480 ESL learners in Malaysia. The participants were learners of 15 years of age. Even though they do not count as Young learners, this study is the only one reporting on using the PVLT as data gathering instrument with not adult learners. The online version of the PVLT is found on Tom Cobb's website: www.lextutor.ca.

2.6. Vocabulary Knowledge Scale

Another vocabulary measure which can serve the purpose of self-assessment is the widely spread Vocabulary Knowledge Scale (VKS) (Paribakht & Wechse, 1999). On the one hand Schmitt (2008, p. 45) praises this type of vocabulary measurement by underlining that it sheds light on what students know, rather than what they do not know, by allowing them to indicate their partial knowledge of a lexical item, it may be more motivating than other types of tests. On the other hand Schmitt (2010, p. 32) criticizes the instrument by claiming that defining depth can be done with extreme difficulty. The format of this test is presented in Table 3.

Table 3. Vocabulary Knowledge Scale

Instruction: Indicate the level you know the word *procrastinate*
1. I don't remember having seen this word before.
2. I have seen this word before, but I don't know what it means.
3. I have seen this word before and I think it means…….
4. I know this word. It means…….
5. I can use this word in a sentence:……..

If a student reports the word is familiar but the meaning is not known, then it is worth no points. This instrument was applied by Lehmann (2009) for the assessment of university students just as Paribakht and Wechse (2006) calibrated the VKS for this age group. However, the VKS has also been designed for Young learners recently. Paribakht and Wechse (2006), Atay and Kurt (2006) and Jóhannsdóttir (2010) used the VKS to assess young learners.

The online version of the test is available on Tom Cobb's website as well. I have no knowledge of any study that has ever used the online VKS yet, however I surmise that applying the online measure would change neither the validity nor the reliability of the test.

2.7. Diagnostic online English and German receptive vocabulary size test for young learners

Most recently a FL vocabulary test has been developed and validated by the researchers of the University of Szeged (Vidákovich, Vgíh, S. Hrebik & Thékes., 2013). The instrument is designed and calibrated to measure diagnostically the vocabulary size of 5th and 6th graders learning English and/or German as a FL. The selection of the target items took place on the basis of frequency lists and corpora and the test is unique in the sense that the words incorporated in the test are almost similar in the two languages. The instrument has a multiple choice test format in that the students see one picture and four words on the screen and they have to decide which words are described by the picture. Contrary to the PPVT, where only one word matches one picture in one task, in this test it might be the case that all four words match the picture or only one word is described by the picture. The test-takers must click on the buttons next to each word and decide whether there is a match or not. The pictures are either simple or complex pictures and students must use the operation of identification or implication to figure out the correct answer. The test requires meaning recognition from the test-takers. The instrument has three versions in both languages. The instrument has always been applied in an online environment on the surface of the eDia platform that has been developed by the ICT specialists of the Education Science Department of the University of Szeged (Molnár, 2013). The test-taking period is short and apart from the test scores background data can be processed virtually immediately after the completion of the data collecting instrument. Table 4 presents one example of the test.

In the pilot study 352 participants took the English test version (Vidákovich et al., 2013). The instrument proved to be robustly reliable and the test versions drew attention to strong relationships and significant correlations with one another. The instrument proved to differentiate well among the test takers. Relevant data were gained concerning the type of words high and low-achieving students know. High-achievers know adjectives and verbs significantly better than low-achievers whereas low-achievers know significantly more nouns than any other word type.

Table 4. Example of a task containing a simple picture

Instruction: Choose words a) – d) that best fit the picture on the left.

a) chair
b) plant
c) table
d) theatre

3. Research

3.1. Purpose of the current study and research questions

In the current study the results of an online test assessing the English as a FL vocabulary of Hungarian 6th graders will be presented. Nikolov and Mihaljevic Djigunovic (2006, p.234) clearly state that the notion 'YL' refers to the age group between early years of studying English up to 14 years of age but not further than that. In this study, the abbreviation 'YL' will sometimes be used.

Even though several diagnostic vocabulary tests exist, they mostly assess learners on paper and pencil. The learners whose vocabulary is investigated in studies published in the past 20 years are adults and in hardly any study have young learners been assessed in terms of word knowledge. Moreover, vocabulary tests either measure receptive or productive knowledge. No existent complex test has been developed and validated.

This means that there is a need to develop a diagnostic online English as a FL vocabulary test assessing young learners' receptive and productive word knowledge. Thus our aim was to develop and validate an online complex vocabulary test for young learners. Considering the general purpose of the research the following research questions were phrased.

(1) Which task of the vocabulary test proves to be the most simple and which proves to be the most difficult?

(2) How do the different items function on the vocabulary test?

(3) How do the young learners perform on the online vocabulary test?

(4) How do the different tasks of the vocabulary test correlate with one another?

(5) How do the high-achievers perform on the productive task of the vocabulary test?

3.2. Methods

Instrument

Up to this point vocabulary had been assessed with tests comprising tasks identical in format. Tests had either assessed receptive or productive word knowledge in one modality. The validity of none of the tests was called into question. However, questions may arise in case an instrument consists of several different tasks. There might be some skepticism whether an item assessed in listening mode would produce the same results as in reading mode. In my view, in an item pool containing 108 words, the overall result achieved in the complex test does not differ from that achieved, say, in a receptive vocabulary test comprising tasks of identical format. According to Melka Teichroew (1982, p. 244) the receptive-productive distinction is rather a continuum than two types of knowledge. It is also asserted that it is not clear where the threshold is found between receptive and productive knowledge (Laufer & Goldstein, 2004). The deficiency in determining the place of this threshold evidences the fact that assessing a number of items in different modalities does not exert an influence on the results.

Besides taking corpus-based data into account, recommendations in the *Hungarian National Core Curriculum* (2007) and Nikolov (2011) were also considered in terms of grouping words based on topics and involving them in the list. The topics suggested were are (1) food and eating; (2) home and furniture; (3) shops and shopping; (4) travelling and transport; (5) jobs; (6) professions and sports. Nikolov (2011, p. 28) suggests 14 broader topics that should be considered by elementary school teachers for classroom practice and she also presumes that the lexis that is included in these topics might be the area of interest for the young language learners. Consequently, I added the most relevant vocabulary of these topics to the list of 2,000 words irrespective of word frequency rank. Magyar and Molnár (2015, p. 48) also support the view of teaching those words to students that they are interested in learning. As a result, my list of words to be assessed was completed by the addition of another 2,000 word families summing it up to 4000 words. This decision is supported by the evidence found by Nation and MacAlister (2010) that the knowledge of the 4,000 most frequent words is the most critical aspect of communicating in a language.

For the measurement tool six of the major topics specified above were selected. There are two reasons for this decision. Not all of the 14 topics could be included in the test and after thorough supervision these six topics included the most frequent vocabulary of all the fourteen. I came to this conclusion after looking at the word lists of these topics and compared them with the frequency lists. Six tasks (Task 1-Task 6) of this complex vocabulary test were intended to assess breadth of vocabulary since most vocabulary tests (Meara, 2009; Nation, 1990; Read, 2000) assess this domain. One task (Task 7) was intended to assess depth of vocabulary. The required word knowledge for task solving was receptive in the first five tasks and in Task 6 and 7 productive word knowledge was the requirement. The VKS was implemented in Task 7. Moreover, I reckoned that it would have been a heavy cognitive load for 6th graders if I had tested depth in more than one task.

The paper-and-pencil version of the vocabulary test was piloted in November 2013 with 103 participants. Item-analysis was conducted to see the functioning of the items and the tasks. With the tools of descriptive statistics results were analyzed and several decisions were made concerning the removal and replacement of items. First of all it was decided that Task 7 would be removed. The main reason for this was that this task showed negative correlations with some of the other tasks. Items with zero standard deviations were also removed and replaced and other instances of replacements occurred in case an item was under .194 (Fauls & Ollé, 2008).After item-analyzing and finalizing the pilot paper-and-pencil vocabulary test, I consulted the information-technology experts of the Institute of Educational Science of the University of Szeged. Assistance was provided by them in converting the finalized paper-and-pencil instruments into an online environment. The test was uploaded onto the online platform developed by Institute of Educational Science called eDia.

In the vocabulary test, all items were designated to three categories. Category 1 words were considered the easiest and Category 3 the most difficult. This classification was determined based on rank, frequency in textbooks used by 6th graders and professional recommendations. Out of the nine items the dispersion of the categories were the following: either four or five Category 1 words, either two or three Category 2 words and either one or two Category 3 words. Category 1 words are normally more frequent grounded on the BNC; however some words related to children's vocabulary with lower ranking were categorized higher than some higher ranked words in the BNC.

Edia is a platform under constant development and is well-suited for efficient data gathering on a large sample. The voice files were also attached to the first two tasks of the vocabulary test. My voice, the researcher's, was recorded reading up the pertaining items. Every task contained a sample task that was presented to the students before they went about taking the test. Taking the vocabulary test took approximately 15 minutes. Students sat down in front of the screen with headsets over their ears so that they could hear the voice file of the first two tasks. The online vocabulary test comprising six tasks to map the English as a foreign language vocabulary of the students. The tasks of the vocabulary test are described in Table 5.

Table 5. Tasks in the diagnostic vocabulary test battery

	Task (Instruction)	**Receptive/ Productive**	**Language skill(s) and modality required Schmitt (2014)**
1	Listen to words and match them with pictures.	Receptive	Listening / Meaning recognition
2	Listen to definitions and match them with words	Receptive	Listening / Form recognition
3	Match 6 written words with 3 pictures	Receptive	Reading / Meaning recognition
4	Match written words with picture	Receptive	Reading / Meaning recognition
5	Match written definitions with words	Receptive	Reading / Form recognition
6	Write word next to picture	Productive	Writing / Form recall

Participants and procedures

The sample was selected by the coordinators of the Institute of Educational Science. The Institute filed a request to schools in Hungary and twelve schools agreed to involve their students in the research. Participants were 282 Hungarian 6th graders. Sampling was non-representative; however this had not been an original goal.

The volunteering schools were given a passcode to be able to log into the eDia platform where the vocabulary test could be accessed. Data were gathered in November 2014 and data processing was performed with the use of the SPSS 17 software.

3.3. Results

As it was described in Chapter 6, the vocabulary test contained 54 items. In all the six different tasks there were eleven items. One item was an exemplary item, one was a distracting item; as a result test-takers had to prove the knowledge of nine item. So, in every task the maximum achievable points were nine making the instrument a 54-point test. Reliability of the test proved to be acceptable (Cronbach's Alpha = .869). In Table 6 the descriptive statistics of the six tasks are presented.

Table 6. Descriptive statistics of the six tasks in the vocabulary test

	Mean	SD	Reliability(Alpha)
Task 1	6.393	2.039	.762
Task 2	3.804	2.534	.812
Task 3	6.135	2.347	.763
Task 4	2.756	2.292	.745
Task 5	2.763	2.293	.770
Task 6	3.380	1.934	.723

Laufer et al. (2004) argue form recognition is expected to be harder than meaning recognition. In the case of the two reading tasks, this argument proved incorrect. In spite of the fact that students performed below 30% in Task 5 (M=2.763), in Task 4 (M=2.756) they achieved even worse refuting the hypothesis that a form recognition task would be more difficult than a meaning recognition task.

Contrary to the paper-and-pencil pilot study that was reported in Chapter 6, on the online test with a larger sample size, participants had the best achievement on Task 1. In the pilot study, Task 3 proved to be the task students which students achieved the best at. Nonetheless students proved to achieve the best on Task 1 and Task 3 during both test procedures. Both tasks are done in meaning recognition modality which is assumed to be the easiest in the hierarchy of modalities (Laufer at el., 2004; Schmitt, 2014). It must also be noted that students scored a lower number of overall test points in the online environment than in the traditional paper-and-pencil environment; however it is not the goal of this study to compare foreign language testing media. Another important finding is that the two reading tasks proved to be the most difficult of all six tasks. Task 4 that required task solving in the modality of meaning recognition and the use of reading skills appeared to be the most difficult for the test-takers whereas in Task 5 demanding task solving in the modality of form recognition and reading definitions and matching them with lexical items participants reached a bit higher number of points than in Task 4, a modality that is assumed to be simpler in the hierarchy. It needs also to be underline that in the task that necessitated the use of productive vocabulary, Task 6 in the modality of form recall, assumedly the most difficult modality, students scored significantly more points than in Task 4 and Task 5. This finding ought to be envisioned in a deeper way. In Task 5 students had to drag a line between the lexical item and the pertaining definition while in Task 6 a set of well recognizable pictures were at their disposal and they had to write one item next to picture. In an online environment it may be easier for students to recall words grounded on recognizing pictures than dragging a line between words and their definitions that might contain lexical items unfamiliar to them. It must not be left out of consideration that the productive task, Task 6, had the lowest reliability value whereas Task 2 in which learners were expected to match definitions they heard to words proved to be the most reliable task.

The histogram clearly shows that the two reading tasks (Task 4 and Task 5) were the most difficult and the first listening task (Task 1) and a reading task in meaning recognition modality (Task 3) were the easiest.

Having examined the six tasks, the descriptive statistics of all the items on the vocabulary test must inevitably be envisioned with particular regard to the item-total correlation values that give account of how each item behaves in a test. In Table 7 the descriptive statistics of the items on the test is presented.

Table 7. Itemwise descriptive statistics of the vocabulary rest

Item	Task	Mean	SD	Item-tot corr.
monkey	1	.706	.456	.338
lion	1	.635	.482	.270
airplane	1	.507	.500	.317
tram	1	.709	.454	.405
swimming	1	.858	.349	.334
helicopter	1	.862	.345	.336
ship	1	.890	.443	.352
camel	1	.858	.232	.426
skating	1	.592	.492	.430
supermarket	2	.585	.493	.386
theatre	2	.862	.345	.404
bake	2	.359	.480	.382
cinema	2	.477	.500	.475
eat	2	.320	.467	.409
hospital	2	.206	.405	.449
learn	2	.253	.435	.406
play	2	.658	.475	.469
sell	2	.534	.499	.420
boat	3	.712	.453	.427
drink	3	.683	.466	.394
drive	3	.676	.468	.486
heavy	3	.737	.441	.431
leg	3	.475	.500	.302
hit	3	.932	.252	.264
pocket	3	.800	.400	.448
quick	3	.682	.466	.513
small	3	.432	.496	.290
busdriver	4	.371	.484	.276
waiter	4	.675	.469	.497
cook	4	.418	.494	.485
firefighter	4	.368	.483	.438
hairdresser	4	.246	.437	.333
mechanic	4	.150	.357	.269
pilot	4	.161	.369	.340
plumber	4	.136	.331	.335
tailor	4	.193	.392	.277
bedroom	5	.676	.471	.204
cook	5	.414	.493	.232
cup	5	.422	.495	.224
curtain	5	.383	.485	.207
dining room	5	.242	.431	.201
open	5	.151	.358	.265
shelf	5	.164	.365	.226
talk	5	.142	.344	.261
wash	5	.181	.387	.282
cake	6	.237	.420	.266

cheese	6	.514	.501	.261
chicken	6	.446	.497	.276
coffee	6	.824	.386	.265
fish	6	.378	.484	.255
hotdog	6	.164	.371	.295
icecream	6	.586	.494	.019
cucumber	6	.192	.314	.332
sausage	6	.162	.364	.288

Frequencies of score ranges

Having analyzed the items in all tasks, the distribution of the score ranges must be envisioned so that a clear picture can be received as far as students' achievement is concerned. Table 42 presents the score ranges and the number of students pertaining to them. Before going into any discussion, it is observable that the test differentiated properly among students with the number high-achievers being more than that of low-achievers.

The maximum point to be received was nine in each of the six tasks, making 54 the overall maximum possible total score. No student achieved 54 points; however twelve reached a remarkable score of 46-48 points. Ten knowledge ranges were determined with five point units except for the top range that was calibrated to the above-mentioned 46-48 since no higher score than 48 was observed. The number of the worst-achieving students, within the range of 1-5 was four and by doing a slight extension to the range of 1-10, the cumulative number of low-achievers is twelve, which is an acceptable number on a sample of 288. This means that not even the 10% of the students achieved below ten points.

By examining the other extremity, the high-achievers, it can be stated that the number of the high-achievers, number of students within the range of 41-48 is 10, which means that not even 5% of the students scored more points than 41. It is inevitable to note that 23 students, almost exactly 10% of the sample scored over 36 points.

As it is expected from a properly differentiating diagnostic test, most students achieved in the range of 40%-60%. The 50% of the total points is 27, which means that in the range of 26-30 points 63 students can be found and 53 students reached the range of 31-35 points. Out of 288 test-takers 116 of them achieved in the average range of 26-35 points, which means that nearly one-third of the sample had an average achievement. Table 8 presents the score ranges of students' achievements.

Table 8. Score ranges of students' achievements

Score range	Number of students
1-5	4
6-10	8
11-15	21
16-20	48
21-25	52
26-30	63
31-35	53
36-40	13
41-45	9
46-48	1

Having analyzed the test score at the item and student levels, it is of paramount importance to examine the correlations among tasks so that deeper relationships can be revealed at task level.

Correlations across tasks in the vocabulary test

The diagnostic instrument assessing word knowledge, as it has been described so far, comprised six tasks. The first two tasks were two listening tasks in meaning and form recognition modality. The third task was a reading task in meaning recognition modality that expected test-takers to match items with

pictures. The fourth and the fifth task were reading tasks in meaning and form recognition modality, respectively whereas the sixth task was a productive writing task in form recall modality. The correlations among these tasks were investigated to see whether the reading tasks had strong relationship with one another and whether the two listening tasks showed any correlations. It was also examined how significantly Task 6 correlated with the rest of the tasks. Table 9 presents the correlation matrix of the six tasks.

Table 9. Correlations among tasks of the vocabulary test

	Task 2	**Task 3**	**Task 4**	**Task 5**	**Task 6**
Task 1	.501**	.434**	.337**	.065	.149*
Task 2		.557**	.530**	.012	.115
Task 3			.517**	.068	.070
Task 4				.368*	.051
Task 5					.476**

**. Correlation is significant at the .01 level (2-tailed).

*. Correlation is significant at the .05 level (2-tailed).

Task 1 and Task 2, the two listening tasks show a significant correlation (r=.501, p<.01), meaning that no matter whether the modality is meaning recognition or form recognition, the two tasks measure the same construct. Task 4 and Task 5 also correlate significantly with a slightly weaker relationship (r=.368, p<.05). Two similar tasks which required the students to match pictures with the items, Task 1 and Task 4 correlate significantly (r=.337, p<.01); however the listening task, Task 2, requiring learners to match items with definitions does not show any correlation with the reading task, Task 5, requiring learners also to match definitions with items. It is intriguing to observe that two related tasks in terms of task solving function have very weak relationship and insignificant correlation with each other within the same test. This result reflects the assumption (Vidákovich et al., 2013) that listening to and reading definitions might be two totally different task solving functions. Furthermore, it is hard to rely on previous research data as young learners' vocabulary had only been assessed in only one modality in each testing instrument. Vocabulary knowledge in different modalities had not been assessed; thus no comparable data are accessible.

By investigating the correlations of Task 6, the productive writing task in form recall modality, crucial information can be procured. Task 6 has a weak relationship with Task 1 but the correlation is significant (r=.149, p<.05). This means that a task requiring the use of a receptive skill, listening, has a stronger relationship with a productive task than with another task also requiring reading skills. Task 6 is also significantly correlated to Task 5. This root of this relationship might be found in the fact that words in these two tasks were ones denoting household items and activities (Task 5) and food (Task 6). These items form a set of words that are usually learned in a collected cluster. The chapters comprising these two sets of words in the course-books used in schools are in close vicinity to each other. Learners that know words meaning food are likely to know those meaning household activities and learners who are not aware of household vocabulary are probably unaware of food vocabulary in a recognition modality, let alone in a form recall modality.

Analysis of variance across tasks of the vocabulary test

For the purpose of looking deeper into the results of the vocabulary test scores, a division was made in the sample and an ANOVA was also implemented. The sample was first divided into three sub-samples on the basis of the test scores. The high-achievers that scored at least two-third of the points, i.e., 36, were classified into first sub-sample. Participants that had a score between 18 and 35 points were classified into the second sub-sample. Finally participants that did not reach at least one-third of the points, i.e., that reached fewer than 18 points were classified into the third sub-sample. The sub-samples formed on the basis of the vocabulary test results are shown in Table 10.

Table 10. The classification of the sub-samples by achievement

Sub-sample	Point range	Number of students
High-achievers	36-54	23
Average achievers	19-35	180
Low-achievers	0-18	79

In Table 45, it is clearly shown that the first, best-achieving, sub-sample scored higher number of points in all tasks than the second and the third sub-sample. In Task 1, 2 and 3, the differences between the high-achievers and the average achievers are striking; however in the last three tasks even the high-achievers of the overall test performed slightly above 50%. It must also be noted that in Task 5 and Task 6 a small gap can be seen between the second and the third, the worst-achieving, sub-sample. The gap is caused by the fact that the low-achievers performed very poorly (M= 2.82 and 3.65 in the two tasks, respectively). It is worth pointing out that both the average and the worst achieving sub-samples performed better in a supposedly more challenging form recall task (Task 6) than in a form recognition reading task (Task 5). This might indicate the fact that poor word knowledge can be more efficiently diagnosed in a reading vocabulary test than in a productive test in form recall modality. One other striking piece of data is that of the low-achievers' task score in Task 4. Almost none of the students in the worst sub-sample could recognize the meaning of any of the words portrayed by pictures. This might be due to poor visual skills or random task solution and careless lining of words to pictures elicited by poor word knowledge. Table 11 presents the descriptive statistics of the three sub-samples.

Table 11. The descriptive statistics of the three sub-samples

	High-achievers **Mean (SD)**	**Average achievers** **Mean (SD)**	**Low-achievers** **Mean (SD)**
Task 1	8.266 (.817)	6.954 (1.548)	4.564 (2.048)
Task 2	7.734 (1.382)	4.183 (2.202)	1.762 (1.512)
Task 3	8.342 (.713)	6.907 (1.712)	3.683 (2.056)
Task 4	5.826 (2.032)	3.072 (2.124)	1.075 (1.214)
Task 5	5.602 (2.344)	2.822 (2.212)	1.752 (1.622)
Task 6	4.781 (1.596)	3.654 (1.868)	2.348 (1.724)

The three sub-samples were compared to see which task result proved to be a determiner in the differences among the students. Having performed the ANOVA, I examined the homogeneity of variances. Firstly, the values on the Levene statistics must be investigated. If the level of significance is less than .05, the post hoc Dunnett-T3 test must be performed whereas in case the level of significance of the Levene statistic is more than .05 then Tukey-B test must be taken (Falus & Ollé, 2008). The levels of significance is presented in Table 12.

Table 12. Levels of significance on the Levene statistic

	Levene Statistic	**Significance**
Task 1	16.802	.000
Task 2	16.091	.000
Task 3	19.242	.000
Task 4	1.824	.160
Task 5	3.743	.021
Task 6	11.142	.000

The Leven Statistic indicates significant differences except for Task 4 and Task 5. The Dunnertt-T3 test was performed in Task, 1, 2, 3 and 6 and the Tukey test was run in case of Task 4 and 5. Besides the Levene statistic the F-values of the analysis of variance were also examined. In each test

significant differences were found amongst the three sub-samples. In Task 1, a high value was found: F(3, 282)=54.77 (p=.000). Task 2 had a lower but significant value: F(3, 282)= 43.90 (p=.000). In Task 3, which the students had the second best achievement among tasks, had the following value: F(3, 282)=51.86 (p=.000). The two most difficult tasks, Task 4 and Task 5, had the lowest F-value of 23.49 and 34.46, respectively. Finally, the productive task, Task 6, had a value of 24.73 (p=.000). In itself, it is not enough to observe the F-values derivative of ANOVA. In cases it was needed Tukey tests were performed (Task 4 and Task 5) to see which task made a significant difference among the sub-sample; at the rest of the tasks, Dunnett-T3 tests were taken. It was found that the tasks results of all the three sub-samples had a significant difference expect for Task 5 and Task 6 where no significance was stipulated between the high-achievers and the average achievers. This means that the four first tasks made the difference between the best achieving sub-sample and sub-sample of students achieving an average number of points.

3.4. Discussion

After analyzing the results of the online vocabulary test, the research questions (RQ) must be answered. The RQs will be listed in order and by referring back to the data analysis in this study the relevant points will be highlighted in answer to research questions.

Task 1, the listening task of meaning recognition modality, proved to be the easiest (M=6.39) and the most difficult task was Task 4, a reading task of meaning recognition modality (M=2.75). It was asserted during data analysis that a task of form recall (Task 6), a supposedly difficult task, proved to be easier (M=3.38). In response to RQ 2, item-total correlation values were evaluated. This value is calculated to see if any of the items do not have responses that vary in line with those items for other tests in the population. In other words, this calculation is performed to check if any item is inconsistent with the averaged behavior of the other items. The minimum of this item-total correlation value, according to the literature, is .194. None of the items, except for 'icecream' (.018) fell below this value. In case a test is under development, it is suggested that the items below .194 should be discarded. In this case there is no possibility to replace 'icecream' so it is not taken out of consideration; however in further research a new item will be implemented in Task 6. Some very low values are come across in, for example, the item the most learners knew, 'hit' had a value of only .264. 'Lion' was also fairly inconsistent with the rest of the test with a value of .270. In an instrument with 54 items, one item not being consistent with the rest of the items might be acceptable. However, it is a striking finding that in Task 5 all of the items' total-correlation values are below .300 but above the .194 limit. Task 5 proved to be the most difficult task as it was stated earlier. Task 5 correlated significantly with Task 4 and Task 6 and had a weak relationship and insignificant correlation with the rest of the tasks. Since none of the items in Task 5 are of unacceptably low item-total correlation values, it can be asserted that Task 5 fits in well with the entire test.

In answer to RQ 3, the sample was divided into score ranges of five point units. Four students fell within the score range of 1-5 points and eight students within the 6-10 point units. This means that twelve students knew fewer than ten words. Even though they had been learning English for two years, at the time of test-taking they had a vocabulary of around ten words. It is incredibly low. As for the average achievers, within the score ranges of 21-25, 26-30 and 31-35, 168 students are found out of the 288 test-takers. By carefully envisioning the badly-achieving, the average-achieving and the well-achieving parts of the sample, a normal distribution can be noticed, which means that the criterion of the classical test theory of proper differentiation is realized. The six tasks showed significant correlations with one another with the exception of Task 5 and Task 6. Task 5 had a weak relationship with Task 2 (r=.012) and a strong relationship but no significant correlation with Task 1 and Task 3 (r=.065 and r=.068, respectively). Task 6 had a weak relationship with Task 2 (r=.115) and no significant correlation with Task 3 and Task 4. It was earlier pointed out in this study that it is hard to find a reason for the near zero relationship between Task 5 and Task 2 because they were of the same modality (form recognition) and the task was the same: matching words with definitions. The only difference was the skills required to solve the tasks: listening and reading. It was supposed that the productive task in form recall modality would be the most difficult task and as such it would be a major differentiating factor among the participants of different word knowledge. As it was discussed earlier in response to RQ 1, Task 6 did not prove to be the most challenging task. However, I intended

to know how high-achievers performed on this particular task to gain better insight into the organization of their vocabulary.

High-achievers had a mean of 4.784 on the productive task, which means that they reached nearly 50% on this task. It is a low value compared to the number of points they reached on Task 1, Task 2 and Task 3. None of them had the maximum nine points on this task and one of the high-achievers on the overall test has as few as two points. This result gives evidence to the fact that a form recall modality task is difficult and most of the Hungarian 6th graders are not prepared to use words or word clusters in production. The classroom implication can be concluded that even learners of good ability must be trained for productive use of the foreign language so that their communicative skills can be improved. Having compared the results of what teachers assumed and what students achieved, it can be asserted that teachers of English of 6th graders generally overestimate the word knowledge of students.

4. Conclusion

The investigation of Young learners' English as a foreign language vocabulary size was a major endeavor since an online data-gathering instruments had to be developed and created. Having conducted a pilot study with the two instruments, item-analysis and several statistical procedures were executed in order that a properly functioning test would be used on large sample for the sake of unveiling correlations and of gaining a deeper insight into the organization of vocabulary.

With regard to the results, the listening task of meaning recognition modality, proved to be the easiest and the most difficult task was Task 4, a reading task of meaning recognition modality. It was asserted during data analysis that a task of form recall (Task 6), a supposedly difficult task, proved to be easier than Task 4 and Task 5. To gain a clear picture of the functioning of the items, total-correlation values were also envisioned. None of the items, except for 'icecream' (.018) fell below a critical value.

Having divided sample was divided into score ranges of five point units, a more sophisticated dataset was gained. Four students fell within the score range of 1-5 points and eight students within the 6-10 point units. This means that twelve students knew fewer than ten words. Even though they had been learning English for two years, at the time of test-taking they had a vocabulary of around ten words. As for the average achievers, within the score ranges of 21-25, 26-30 and 31-35, 168 students are found out of the 288 test-takers. By carefully envisioning the badly-achieving, the average-achieving and the well-achieving parts of the sample, a normal distribution can be noticed, which means that the criterion of the classical test theory of proper differentiation is realized.

References

Abduallah, K.I., Puteh, F., Azizan, A.R., Hamdan, N.N., & Saude, S. (2013). Validation of a controlled productive vocabulary levels test below the 2000-word level. *System, 21*(2), 352–364.

Alderson, J.C. & Huhta, A. (2005). The development of a suite computer-based diagnostic test based on the Common European Framework. *Language Testing, 22*(3), 301-320.

Augustin Llach, P.M. (2011). *Lexical errors and accuracy in foreign language writing. Second Language Acquisition*. Bristol: Multilingual Matters.

Atay, D. & Kurt, G. (2006). Elementary school EFL learners' vocabulary learning: The effects of post-reading activities. *The Canadian Modern Language Review, 63*(2), 255-273.

Bachman, L.F. (1990). *Fundamental considerations in language testing. Oxford:* Oxford University Press.

Bogaards, P. (2000). Testing L2 vocabulary knowledge at a high level: the case of the Euralex French Tests. *Studies in Second Language Acquisition 21*(2), 128-160.

Catalan Jimenez, R.M., & Terrazas Gallego, M. (2008). The receptive vocabulary of English foreign language young learners. *Journal of English Studies 5*(1), 173–191.

Chostelidou, D., Griva, E., Ioannidis, T., & Panitsidou, E. (2012). Multilingual learning for specific surposes: Identifying language strategies, awareness and preferences. *Procedia 46*, 1419-1423.

Ellis, N.C., & Beaton, A. (1993). Psycholinguistic determinants of foreign language vocabulary learning. *Language Learning 43*(4), 559-617.

Fitzpatrick, T., Al-Qarni, I., & Meara, P. (2008). Intensive vocabulary learning: a case study. *Language Learning Journal*, *36*(2), 239-248.

Hungarian National Core Curriculum (Magyar Nemzeti Alaptanterv) (2007). Oktatási Minisztérium. Budapest.

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

Jackson, H., & Zé Amvela, E. (2011). *Words, meaning, vocabulary. An introduction to English lexicology.* London: Bloomsbury.

Jóhannsdóttir, R. (2010). English in the 4th grade in Iceland. Exploring exposure and measuring vocabulary size of 4th grade students, *Menntakvika 1*(1), 1-20.

Kilgarriff, A. (1997). Putting frequencies in the dictionary. *International Journal of Lexicography, 10*(2), 135–155.

Laufer, B. (2001). Quantitative evaluation of vocabulary: what it is good for and how it can be done. In C. Elder, A. Brown, E. Grove, K. Hill, N. Iwashita, T. Lumley, T. McNamara, K. O'Loughlin (Eds.), *Experimenting with Uncertainty*. (pp. 241-250). Cambridge: Cambridge University Press.

Laufer, B. & Nation, I.S.P. (1999). A vocabulary size test of controlled productive ability. *Language Testing*, *16*(1), 33-51.

Laufer, B., Elder, C., Hill, K., & Congdon, P. (2004). Size and strength: do we need both to measure vocabulary knowledge? *Language Testing 21*(2), 202–226.

Lehmann, M. (2009). *Assessing English majors' vocabulary at the University of Pécs*. Unpublished PhD thesis. University of Pécs.

Lewis, M. (1993). *The lexical approach.* Cambridge: Cambridge University Press.

Matsuoka, W. & Hirsh, D. (2010). Vocabulary learning through reading: Does an ELT course book provide good opportunities? *Reading in a Foreign Language 22*(1), 56–70.

Meara, P. (2009). *Connected words*. Amsterdam: John Benjamins Publishing.

Milton, J. & Fitzpatrick, T. (2014). *Dimensions of vocabulary knowledge*. Basingstoke: Palgrave Macmillan.

Molnár, Gy. (2013). Számítógépes játékon alapuló képességfejlesztés: egy pilot vizsgálat eredményei. *Iskolakultúra*, *21*(4), 3-12.

Nagy, J. (2004). A szóolvasó készség fejlődésének kritériumorientált diagnosztikus feltérképezése. *Magyar Pedagógia*, *104*(2), 123–142.

Nagy, W., Anderson, R. & Hermann, P. (1987). Learning word meanings from context during normal reading. *American Educational Research Journal*, *24*(2), 237-270.

Nagy, W. & Scott, J.A. (2000). Vocabulary processes. In M.L. Kamil, P.B. Mosenthal, P.D. pearson & R. Barr (Eds.), *Handbook of Reading Research* (pp. 269-284). Mawah, NJ: Erlbaum.

Nassaji, H. (2003). L2 vocabulary learning from context: strategies, knowledge sources, and their relationship with success in L2 lexical inferencing. *TESOL Quarterly*, *37*(4), 645-670.

Nation, I.S.P. (1990). *Teaching and learning vocabulary.* Boston: Heinle and Heinle.

Nation, I.S.P. (1999). Fluency and accuracy. In B. Spolsky (Ed.), *Concise Encyclopaedia of Educational Linguistics*. (pp. 611-628). Oxford: Elsevier Science.

Nation, I.S.P. (2001). *Learning vocabulary in another language.* Cambridge: Cambridge University Press.

Nation, I.S.P. (2011). Vocabulary research into practice. *Language Teaching*, *44*(4), 529-539.

Nation, I.S.P. (2013). *Learning vocabulary in another language*. Second edition. Cambridge: Cambridge University Press.

Nation, I.S.P. & Beglar, D. (2007). A vocabulary size test. *The Language Teacher, 31*(7), 9–13.

Nation, I.S.P. & Macalister, J. (2010). *Language Curriculum Design*. New York: Routledge.

Nikolov, M. & Mihaljevic Djigunovic (2006). Recent research on age, second language acquisition, and early foreign language learning. *Annual Review of Applied Linguistics, 26*, 234-260.

Nikolov, M., & Szabó, G. (2011). Az angol nyelvtudás diagnosztikus mérésének és fejlesztésének lehetőségei az általános iskola 1-6. évfolyamán [Possibilities of developing English diagnostic tests for years 1-6 in the primary school]. In B. Csapó and A. Zsolnai (Eds.), *A kognitív és affektív fejlődés diagnosztikus mérése az iskola kezdő szakaszában.* (pp. 13-40). Budapest: Nemzeti Tankönyvkiadó. 13-40.

Paradis, M. (2004). A neurolinguistics theory of bilingualism. Amsterdam: John Benjamins.

Paribakht, T.S., & Wechse, M. (1999). Reading and incidental L2 vocabulary acquisition. An introspective study of lexical inferencing. *Studies in Second Language Acquisition, 21*(2), 195-224.

Paribakht, T.S.**,** & M. Wesche. (2006). Lexical inferencing in L1 and L2: Implications for vocabulary instruction and learning at advanced levels. In H. Byrnes, D. Weger-Guntharp, & K. A. Sprang (Eds.), *Educating for Advanced Foreign Language Capacities: Constructs, Curriculum, Instruction, Assessment. (pp.* 118-135). Washington, DC: Georgetown University Press.

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

Schmitt, N. (2008). Instructed Second Language Vocabulary Learning. *Language Teaching Research 12*(3), 329-363.

Schmitt, N. (2010). *Researching Vocabulary: A Vocabulary Research Manual*. New York: Palgrave Press.

Schmitt, N. (2014). Size and depth of vocabulary knowledge: What the research shows. *Language Learning, 64*(4), 913–951.

Schmitt, N., & Schmitt, D. (2014). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching, 47*(4), 484 - 503.

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behavior of two new versions of the Vocabulary Levels Test. *Language Testing 18(1),* 55–88.

Schmitt, N., & Schmitt, D. (2014). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching, 47*(4), 484 - 503.

Schoonen, R. & Verhallen, M. Kennis van woorden. De toetsing van diepe woordkennis. *Pedagogische Studiën 75*(3), 153-168.

Scott, V., & de la Fuente, M.J. (2008). What's the Problem? L2 Learners' use of the L1 during consciousness-raising, form-focused tasks. *The Modern Language Journal 92*(1), 100-113.

Thékes, I (2014). Egy kontrollcsoportos angol szótanulási kísérlet eredményei magyar főiskolások körében. (Results of an English as a foreign language vocabulary learning control-group experiment amongst college students) In: J. Bárdos, L. Kis-Tóth, & R. Racskó (Eds.), *Új kutatások a neveléstudományokban. Változó életformák, régi és új tanulási környezetek. (New studies in education science. Changing lifestyles, old and new learning environments)* Budapest: MTA Pedagógiai Tudományos Bizottság, 269-181.

Thornbury, S. (2002). *How to teach vocabulary*. London: Pearson.

Vidákovich T., Vígh T., Sominé Hrebik O., & Thékes I. (2013). Az angol és német nyelvi szókincs online diagnosztikus tesztelése a 6. évfolyamon. (Diagnostic assessment of English and German as a foreign language vocabulary amongst 6th graders). *Iskolakultúra 23*(11), 117-131.

Webb, S., & Sasao, Y. (2013). New directions in vocabulary testing. *RELC Journal, 44*(3), 263 - 278.

Xue, G., & Nation, I.S.P. (1984). A university word list. *Language Learning and Communication 3*(2), 215-229.

Zareva, A. (2007). Structure of the L2 mental lexicon: How does it compare to native speakers' lexical organization? *Second Language Research, 23*(2), 123-153.

Author

Istvan Thekes, Doctoral School of Education, University of Szeged, Szeged (Hungary). E-mail: jerry@jerrythekes.com.