Second Language Vocabulary Acquisition: Spacing and Frequency of Rehearsals

by

Gerlinde Weimer-Stuckmann
B.A., Concordia University, 1990

A Thesis Submitted in Partial Fulfillment
of the Requirements for the Degree of

MASTER OF ARTS

in the Department of Germanic and Slavic Studies

Second Language Vocabulary Acquisition: Spacing and Frequency of Rehearsals

By

Gerlinde Weimer-Stuckmann

B.A., Concordia University, 1990

Supervisory Committee

Dr. Ulf Schuetze, Supervisor
(Department of Germanic and Slavic Studies)

Dr. Catherine Caws, Co-Supervisor
(Department of French)

Dr. Li-Shih Huang, Outside Member
(Department of Linguistics)

Supervisory Committee

Dr. Ulf Schuetze, Supervisor
(Department of Germanic and Slavic Studies)

Dr. Catherine Caws, Co-Supervisor
(Department of French)

Dr. Li-Shih Huang, Outside Member
(Department of Linguistics)

ABSTRACT

The theories of spaced rehearsal have established recurring encounters as a key aspect in vocabulary retention. However, how and how often these review sessions should be scheduled is still a controversial debate. This study reports on a large-scale study at the University of Victoria in the fall of 2008 on Second Language Vocabulary Acquisition (SLVA). Over a period of 13 weeks 117 students practiced 200 German lexical items using a multimodal vocabulary learning web application and research tool ViVo. First, this study contrasted the rehearsal conditions of graduated intervals and uniform spaced intervals studied in 5 practice sessions. Second, it contrasted frequency test conditions. Students who had practiced 2 or 3 times were compared to students who had practiced 4 or 5 times. Results showed no significant difference between uniform interval spacing and graduated interval spacing even though students studying on a uniform practice schedule demonstrated slightly higher test results. With regard to frequency, students practicing 4 or 5 times significantly outperformed those students studying only 2 or 3 times.

Table of Contents

THIS PAGE MISSING FROM ORIGINAL DOCUMENT SUBMITTED

List of Tables

List of Figures

Acknowledgments

I would like to thank my supervisor Dr. Ulf Schütze for the help and expertise he has provided during research and writing of this thesis. This was an inspiring project and a learning journey. Furthermore, I wish to thank participating students and the instructors in the pilot study and research: Dr. Matt Pollard, Dr. Helga Thorson, Dr. Charlotte Schallié, Elena Trebes and Kathrin Siedentopf for their cooperation and assistance. This project could not have taken place without them. I would also like to thank Dr. Catherine Caws for her feedback as the second reader.

Finally, and most importantly, I would like to thank Thomas Weimer for the many hours of programming and technical support spent on creating and implementing ViVo© as a research and learning tool. This project was quite an endeavour he shared and supported.

Dedication

Für Thomas, Anne-Madeleine und Timo

1. Introduction

Our success in conveying our thoughts and ideas in a second language (L2) depends on the development of *"robust"*[1]corpora. Words are a language's building blocks. We need words. But –

What is a word?

What does it entail *to know* a word?

How are words represented in our minds?

How do we process them?

How do we retain them?

How do we retrieve them?

These issues have been predominant in Second Language Vocabulary Acquisition (SLVA) research. If the development of robust corpora is the objective, what then constitutes best pedagogical practice? Contributing factors of SLVA are abundant; conditions are interrelated, and most agree that the magnitude of the learning task is daunting. But, as Horst and Cobb (2006) criticized, even though present day research has arrived at a number of useful insights on behalf of best practices for SLVA, these have had little impact on the pedagogical implementation in Canada's school language programs. The programs are reluctant to include explicit vocabulary practice.

> "There is a distinct reluctance by educators trained in the communicative paradigm to implement decontextualized word-learning activities such as the study of word and definition pairs - even though experiments such as Laufer's (...) consistently point to their effectiveness." (Horst & Cobb, 2006, p.5)

[1] "Robust learning" is a term defined by Van Lehn as a. long-lasting b. transferable and c. as learning that promotes future learning (Van Lehn, 2006, p.6).

This research focuses on two specific SLVA aspects within the context of decontextualized practice in a computer assisted language learning setting (CALL): the spacing effect for rehearsals and the frequency of these rehearsals.

The concept of *spaced learning*, also referred to as *cyclical learning,* is to rehearse a learning task after some time has passed. In particular, this study explored two types of rehearsal schedules: graduated and uniform. Karpicke and Roediger (2007b) pointed out, that it is astonishing that graduated practice schedules are considered to be best practice in pedagogy even though there is not a large base of research with consistent evidence to back this claim. The situation is a similar one regarding rehearsal frequency. Learners are told that they need to practice vocabulary more than once, but recommendations based on research findings, which define what would constitute a good practice frequency, vary considerably.

The emergence of information and communication technologies have changed and enriched language learning. Multimedia CALL settings provide an environment for explicit study and reinforcement of vocabulary in systematically spaced practices. This study was conducted using the web-based application ViVo© (henceforth referred to as ViVo), a software that was designed and programmed as a research and learning tool. It allowed for an enriched multimedia environment and enhanced learning by facilitating the intense processing of the target items. At the same time, the software was programmed to implement the researched practice schedules. Hence, it allowed for a field research conducted in an educational environment.

In recent years, studies in cognitive psychology have explored repetition and spacing (i.e., Baddeley, 1990; Bahrick, P., Bahrick, H., Bahrick, L., & Bahrick, A., 1993; Balota, Duchek, & Logan, 2007; Bloom & Shuell, 1981; Carpenter & De Losch, 2005; Cull, 2000; Dempster, 1987; Nation, 2001). Some

research studies have looked into frequency of exposure (i.e., Cowan, 2001, 2005; Hulstijn, 2002, 2003). However, as Balota et al. (2007) have pointed out, not much research has explicitly examined and compared uniform spaced interval schedules and graduated interval schedules within the context of SLVA or addressed the factor of frequency. Yet, these issues are, as many would agree, key concepts of SLVA.

This research researched the optimal spacing of practice sessions and the optimal number of rehearsals in a CALL setting. It thereby aimed to contribute to our understanding which spacing schedule and which rehearsal frequency would benefit students' vocabulary retention most.

2. Background

Second language research on vocabulary retention crosses disciplines of neurology, linguistics, psychology, applied linguistics, and education. The following sections will therefore also address approaches and research findings in these disciplines that are relevant to SLVA and the research focus. They are organized hierarchically from a broad interdisciplinary perspective to the specific research on SLV spacing effects.

The first section presents issues of variation in terminology. Key terms are lexemes, lemmas, lexical units, vocabulary items, SLVA, second language (L2), first language (L1), and depth or breadth of knowledge.

The second section addresses research on memory and retention. Key terms are functional approach, structuralistic approach, and multi-store memory concepts. It then focuses on theories of language processing in the mental lexicon with the key concepts of symbolist and connectionist approaches.

With SLVA in mind, the third section looks at memorization systems. Spaced learning concepts are explored in more depth. Key terms are the forgetting curve, cyclical learning, distributed practice versus massed practice, uniform spaced practice, and graduated interval recall –also referred to as expanded retrieval.

In the fourth section, a brief outline illustrates how these theories have translated into educational practices. It concludes the need for further research and presents the research questions of this thesis.

2.1. Terms and definitions

Catering to the development of robust L2 corpora may be the set goal of SLVA, but variations in terminology have greatly hindered the task of defining

what it is we wish to acquire (Singleton, 1999). The following will therefore briefly describe how the terms *words* and *L2* are used in this research.

2.1.1 Words – lemmas- lexemes- vocabulary items

In order to learn language we need words. Yet, when asked to describe the properties of a *word*, SLA researchers' definitions are diverse. Schmitt (2000) regarded the term *word* as too general because it does not take morphological permutations into consideration. He argued for the term *lexical unit* or *lexeme* as it allows for multi-word phrases as a single meaning unit. Yet, he chose to continue to use *word* until more a precise term had been coined. [2] Crystal (1987) defined a *lexeme* as the smallest contrastive unit in a semantic system that is listed as separate entry in dictionaries. Aitchison (2003) differentiated between a representation in the mental lexicon of word form and word meaning. She referred to the latter as *lemmas*: "The mental lexicon contains whole words. These are likened to coins with lemmas (meaning and word class) on one side and word forms (sounds) on the other (p. 249)". Nation (2001), however, defined a *lemma* as a base word and its inflections. Singleton (1999) discussed various approaches to define *word* taking for instance their orthographic representation, their acoustic characteristics, their morphosyntactic levels, and their lexicogrammatical information into account. Nation (1990, 2001) listed different aspects of word knowledge (meaning, written form, spoken form, grammatical behaviour, collocations, register, associations, and frequency). This list is often cited as reference of what we know of a word.

Read (2004) then expanded these aspects of word knowledge. He discussed them as aspects of depth and breadth of knowledge[3] and introduced

---

[2] See Schmitt (2000) for a comprehensive overview of the terminology used.

[3] See also Anderson and Freebody (1981); Nation (2001) and Schmitt (2000) for the discussion on depth and breadth of word knowledge.

the terms "precision of meaning, comprehensive word knowledge and network knowledge" (p.231). In this, he differed from Anderson and Freebody's (1981) definitions that defined breadth as the quantity of words learned. Anderson and Freebody have contrasted the size of a learner's corpora to the quality and depth of knowledge[4]. Read further argued that depth of knowledge reflected social and cultural background, occupation, personal interests, and education as well as vocabulary knowledge. The acquisition of these aspects was regarded as a highly individual continuous process: parallel or sequential, with varying degrees of proficiency, and taking place at varying times.

In this context, many researchers (i.e., Nassaji, 2004; Nation, 2001; Paribakht & Wesche, 1999; Read, 2004; Schmitt, 2000) had come to think of vocabulary acquisition as a process. Assessment of *depth* or *proficiency* in any of these acquisition aspects was therefore addressing performance at a particular point on a procedural continuum. According to Scherfer (1994), vocabulary retention was therefore flexible and fluid: a dynamic interaction of storage, retrieval, and processing.

So, what does it mean for this research project to know a word, a lexical item, a lemma? When citing research of others, I used the term the author used. For the purpose of the present research, I chose to use *vocabulary item* for pragmatic reasons. It included the following lexical categories: a. a single item (i.e., *heute* – today), b. phrases (i.e., *Auf Wiedersehen* – goodbye, *einkaufen gehen* – to go shopping), and c. single items with the grammatical gender (i.e., *die Lampe* – lamp). The choice of these items was based on the text book entries of Deutsch NaKlar[5]. Most importantly, at this beginner's stage of acquisition, the term *vocabulary item* referred to an assumed and admittedly simplistic 1:1

[4] This distinction has been discussed controversially (Read, 2004, p. 204).

[5] See a more detailed description of *Deutsch NaKlar's* corpora in 3.2.3.

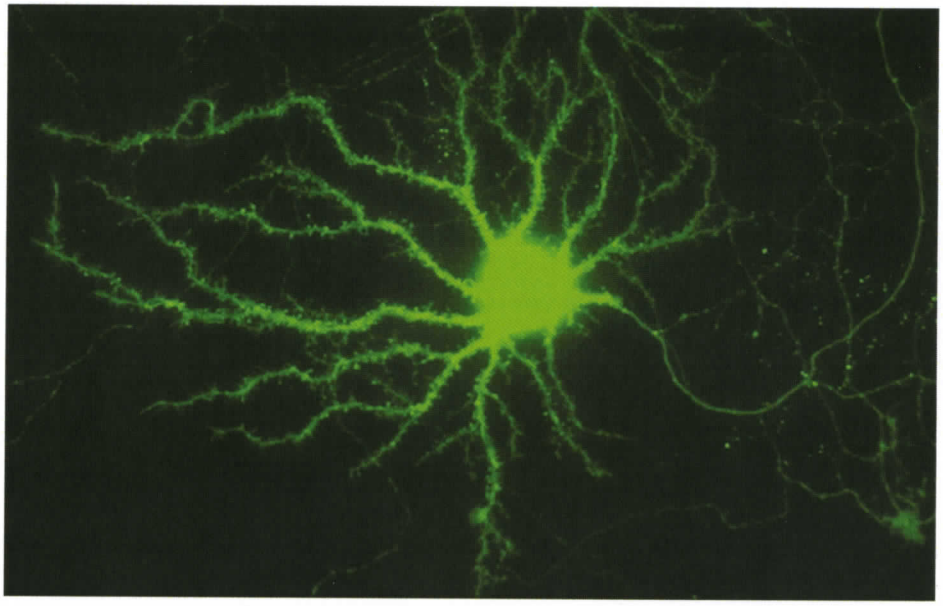
relation of (L1) and (L2) concept as presented in the textbook corpora. I further assumed that the vocabulary items used in this research were positioned at the beginning of the procedural continuum because the participants of this study did not have prior knowledge of the target language.

2.1.2 Second language – foreign language

This study took place in a foreign language context. Participants did not use the target language German outside the course environment nor were they immersed in other learning tasks that required them to use German. Students who had indicated their extensive use of the target language German with friends or family members were not included in the data (see 3.3.1 and 3.4.2 for details). Typically, students who are immersed in a second language environment encounter the target language outside class, too (i.e., Anglophone students learning French in Québec), or they use it to complete other assignments (i.e., Anglophone students learning Geography in French as part of their French immersion program). However, the term Second Language Vocabulary Acquisition (SLVA) is well-established in research. I therefore used this term to describe the vocabulary acquisition of this study's target language. And, for simplicity, I referred to the target language as L2 even though German was the third foreign language for some students.

2.2 Memory

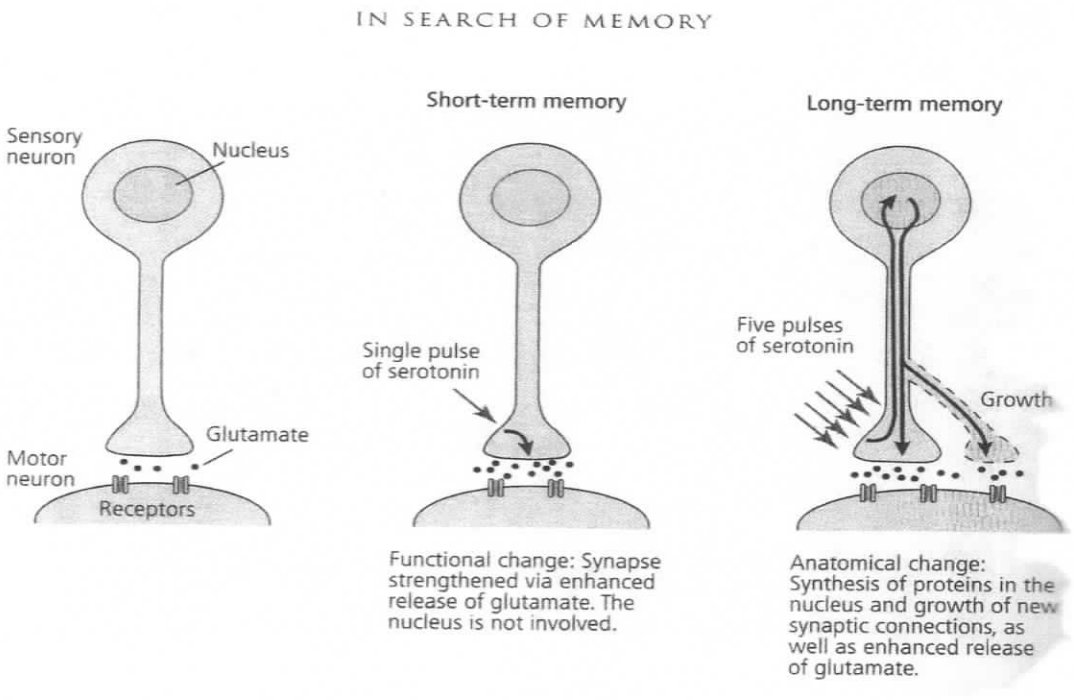
Figure 1: Nerve cell as shown in the Max-Planck Gesellschaft press release (January 17, 2006)



Language is a feat of our brain. Therefore, second language acquisition in all its aspects: input, processing, storage, retrieval, output is inseparable from cognitive processes of our brain.

The notion that memory is comprised of different components has been current for decades (Baddeley, 1986, 1990, 1997, 2007). This notion was further reinforced by neurophysiological theories that identified structural differences between short-term and long-term memory (Başar, 2004). Kandel (2006) described short term memory as a functional exchange of neurotransmitters and long term memory as an anatomical change involving the synthesis of protein structures as well as the growth of new synaptic connections.

Figure 2: Structural changes in short term and long term memory as shown in Kandel (2006, p. 256)



Memory is a procedure (Başar, 2004) with processes of encoding, retention, and retrieval. Encoding can be seen as the first encounter processing sensory data. The term retention refers to the data retained over a period of time that can span seconds to years. The retrieval process makes this data accessible. In milliseconds, (Friederici, 2002; Kandel, 2006) we can do both: access memories dating from years ago or experience an immediate recall of something encountered mere seconds ago. Conway, Jarrold, Kane, Miyake, & Towse (2007) therefore described memory as "the ability to mentally maintain information in an active and readily accessible state, while concurrently and selectively processing new information" (p. 3).

2.2.1. Structuralistic approach and functional approach

Models of memory at work can be related to two strongly debated theoretical schools of research: the structuralistic approach and the functional

approach. In the first, memory is comprised of different components. In the latter, retention differences are viewed as differences of depth of processing.

Two of the most prominent theories emerging from the structuralistic approach were the multi-store models of Atkinson and Shiffrin (1968) and cognitive psychologists Baddeley and Hitch (1974). In 1968, Atkinson and Shiffrin introduced their *two-store model* of memory that distinguished between short-term memory and long-term memory. In the following years, they included a sensory buffer zone in their model (see below). This model was highly debated and modified by other researchers many times.[6]

The *three-store model* introduced by Baddeley and Hitch (1974) elaborated more on the short-term memory. Baddeley and Hitch introduced the concept of a *working memory.* Their model is described in more detail below because its findings were often cited when addressing SLA conditions. This model consists of the registers sensory buffer, working memory and long-term memory. These registers each serve a different purpose and differ in capacity, length of retention, and encoding processes.[7]

The sensory buffer is modality based (i.e., phonological, visual, and olfactory). It receives sensory data input but does not forward it for further processing unless this input receives more attention. Data is only then remembered when it is attended to.

Baddeley and Hitch (1974) have researched the short-term memory component more closely and refer to it as the *working memory*. They identified its processes that enable the transfer from short-term memory into long-term memory. They described the working memory as a construct with three

---

[6] See Baddeley (1986) for a historical overview of this controversial debate.

[7] Various approaches differ in the detailed description of these registers. This categorization is therefore only general. Başar (2004) presented a more detailed description of these approaches.

components: the central executive and its two substructures, the visuo-spatial sketchpad, and the phonological loop[8]. The central executive controls incoming sensory data and manages its processing. The visuo-spatial sketchpad component processes nonverbal information, whereas the phonological loop stores phonological information and rehearses it on a subvocal level. This aspect is particularly interesting for SLVA because Baddeley and Hitch (1974) argued that graphic presentations must first be decoded to a phonological representation in order to enter this part of the working memory. Otherwise, forgetting would occur within seconds. This advocates for SLVA strategies that incorporate a strong phonological component (Ellis and Beaton, 1993).

Long-term memory comprises a human being's world knowledge, emotions, and thoughts. Its workings are believed to trigger the formation of a complex system of neurophysiological structures.

Baddeley and Hitch's (1974) model has been embraced and modified by many researchers. The current psycholinguistic concept of the working memory places it as part of the long-term memory (Cowan, 1996; Levelt, 1993). However, this model has also sparked a controversial debate.

Because multi-store models did not seem to account for all aspects of retention, other models were introduced. Parallel to Baddeley and Hitch's studies, Craik and Lockhardt (1972) developed a concept that seemed to fill this gap: a theory of *levels of processing*, a functional approach.

In contrast to the multi-store models, a model of the functional approach only displays one store. Retention differences are viewed as differences of depth of processing. One of the most prominent models by Craik and Lockhardt (1972) is briefly outlined below.

---

[8] In 2000, Baddeley added the episodic buffer as a fourth component. He described this component as the "interface" and a "binding mechanism" between the three working subsystems and the long-term memory (Baddeley, 2007, p. 13).

According to Craik and Lockhardt's (1972) theory of *levels of processing*, successful retention depended on the depth of data encoding. This theory therefore allowed for the concept of a continuum of processing. The definitions of the storage registers short-term and long-term were of less importance. Input was graded from low-level importance (i.e., sound, colour, and rhythm) to high-level input processing (i.e., semantic analysis). Furthermore, in depth processes varied according to factors such as motivation, perceived importance, effort, and conceptual familiarity.

Yet, again this approach was not able to cover all aspects of the memory phenomenon.[9] It could not account for long-term retention of data that had only been encountered on a seemingly low level. For example, we might remember a melody that we have heard only once. Furthermore, this theory was criticized because it did not define what constituted every level of processing (Hulstijn, 2003; Baddeley, 2007).

2.2.2 Relevance for SLVA

Theories of a structuralistic approach and of a functional approach have informed many research projects on SLVA. With regard to the sensory buffer and the need to attain to its incoming stimuli, Schmidt (1994, 2001) pointed out the necessity to draw attention to the learning task: "It is argued that unattended stimuli persist (...) for only a few seconds at best, and attention is the necessary and sufficient condition for long-term memory to occur" (Schmidt, 1994, p. 16). The theory on levels of processing was of interest to the concept of "depth of knowledge" (i.e., Henriksen, 1999; Nation, 2001; Read, 2004). Stimuli that had been attended to on more than one level and more than one mode were

[9] For a research overview of this controversial debate see Baddeley, 1986, p. 26.

retained better. Finally, Doughty (2001) pointed out three pedagogical implications that can be closely related to the theory of memory and retention: noticing, processing and encoding. In particular, the timing aspect is of interest to this thesis. Is there a sensitive time, in cognitive terms, for pedagogical intervention in the language acquisition process? And, can optimized rehearsal patterns make this acquisition more time-efficient?

Transfer of SLV into the long-term memory is the goal of SLVA. Yet, our long-term memory is not an infallible sum of everything we have encountered and memorized. We forget.

2.2.3 Forgetting

Various approaches have aimed at explaining the phenomenon of forgetting. Tombaugh and Hubley (2001) have reviewed recent research and suggested a classification into two controversy concepts. One discusses retention as an issue of encoding and storage, the other focuses on retrieval processes.

According to the first theory, we forget because we have not been reminded. Items in our memory must be revisited. Otherwise, they are lost. They only continue to exist because we access them regularly and thereby ensure that they remain active in our minds.

Another approach (Arbinger, 1984) focuses on the retrieval aspect. Memories are not lost, but their access route is not available. In this context, Aitchison (2003) referred to the tip of the tongue phenomenon. People feel that they know the word and claim that *it is on the tip of their tongue* but they cannot access it readily.

However, neither approach has been able to explain all facets of memorization. For example, we may remember something we have encountered only once. This then contradicts the encoding-storage theory.

In conclusion, there are different approaches to describe the ways how the human memory works. Cognitive psychology research has informed our understanding of encoding, storage and retrieval of stimuli. How these findings translate into models and concepts relevant to L1 and L2 acquisition is discussed in the following section on the mental lexicon.

2.2.4 Mental lexicon

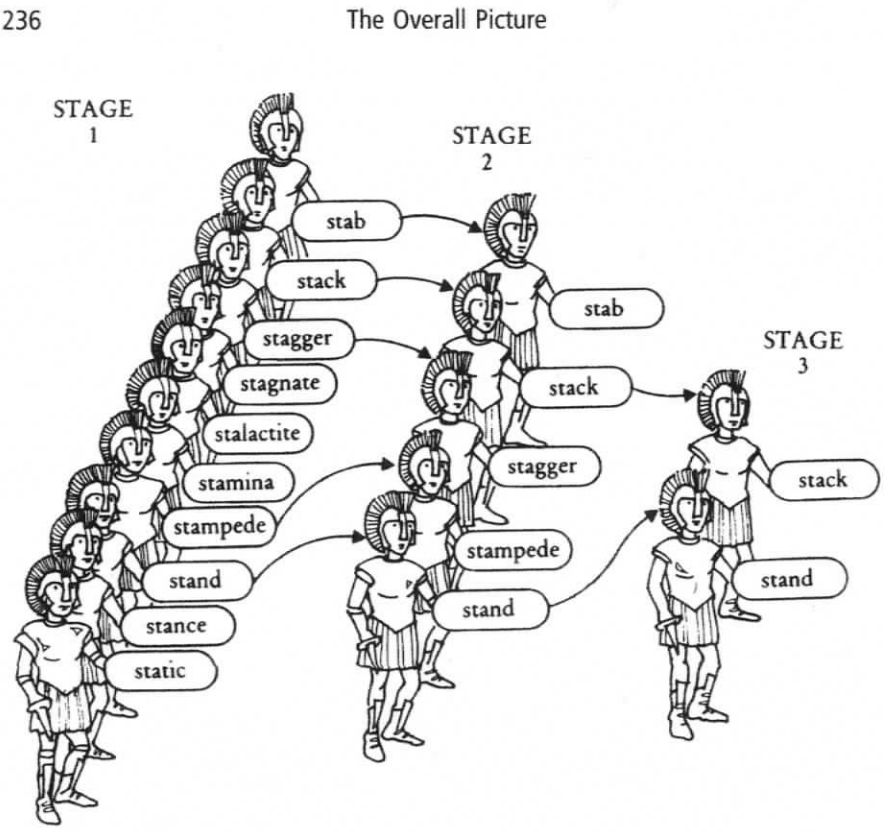
Language requires two procedures: recognition as part of comprehension, and retrieval as part of production. Aitchison (2003) stated that "the human word-store" (p.10) is often referred to as mental lexicon. *λεξικό - lexicon* is the Greek word for dictionary. The term *lexicon*, however, as she cautioned, is misleading because it suggests a conceptual similarity between dictionary organization (i.e., compiling and storing items in a hierarchical order) and humans' mental 'dictionary'. Even today, the more commonly used computer metaphor refers to the concept of an organization where lexical units are *stored* and accessible for *retrieval* following algorithms similar to computer organization.

How these processes of retention and retrieval manifest themselves is still a controversial debate between hierarchically based models such as Marslen-Wilson and Welsh's (1978) cohort model and connectionist approaches such as the interactive-activation model.

Marslen-Wilson and Welsh (1978) developed a model based on the activation of auditory word detectors. In a staggered matching process

mismatches were progressively removed from the set until the final target item was identified.

Figure 3: The cohort model as shown in Aitchison (2003, p. 236)

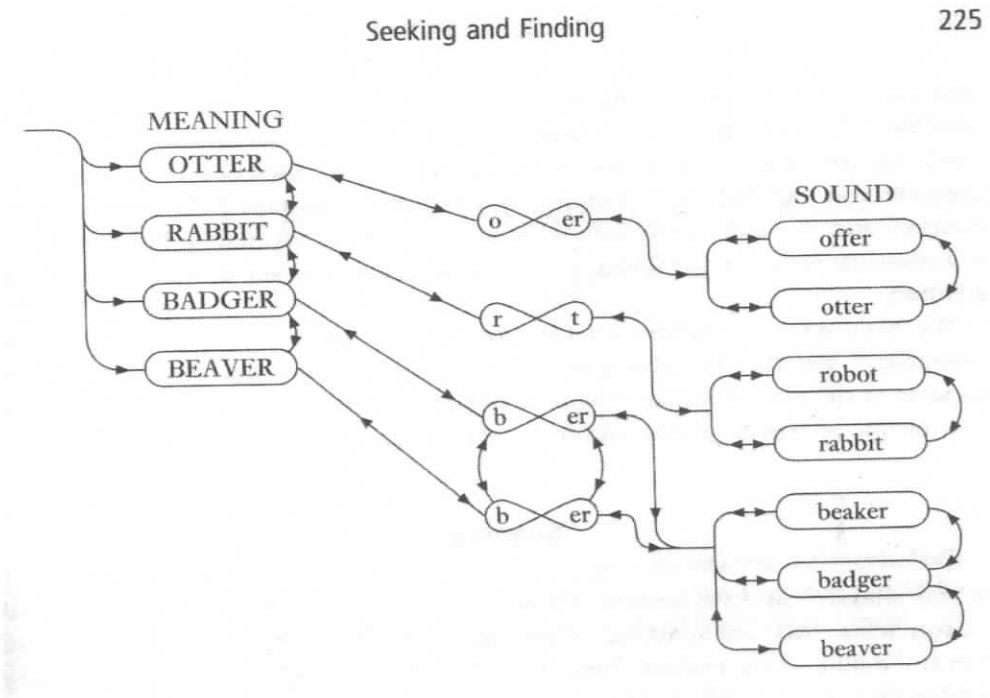


Subsequent variations of this model took into account that context would also contribute to the selection process. The acoustic bottom-up approach based on the binary function *true or false* was expanded to accommodate context choices: would this word make sense in this environment? Aitchison (2003) and Singleton (1999) pointed out that this variation came close to the concept of interactive activation described below.

Like the cohort model, the interactive activation model activates more items than needed. However, the selection process differs because semantics influence it. This processing has been compared to a condition where electrical impulses spark more or less energy for different items. Depending on the energy

level of these activations, the most activated items are then chosen. In this, it echoes what we know of the neural activity in the brain (Aitchison, 2003, p. 237).

Figure 4: The interactive activation model as shown in Aitchison (2003, p. 225)



Processing is therefore not hierarchical or bottom-up, but linked within a network.

In conclusion, Aitchison (2003) stated that though the mental lexicon concepts are discussed controversially most researchers now agree that the lexicon contains whole words that are organized in a fluid network. Furthermore, words are comprised of two parts: meaning- word class and sound structure. Next, the word selection is determined by the activation-inhibition duality while simultaneously allowing for an interactive parallel processing.

The question is: how does our knowledge of mental lexicon components translate into SLVA curriculum design? There is a perceivable gap between theoretical concepts and SLVA objectives.

Schmitt (2000) argued that in the past many SLA approaches did not really deal with vocabulary, "with most relying on bilingual word lists or hoping it would just be absorbed naturally" (p. 15). In her research, Laufer (2006) argued that the "default hypothesis" (p. 152) in vocabulary learning, absorbing through exposure, was still common practice. Yet, as she explained, it could not lead to promising results because:

1. Learners did not focus on unfamiliar words when attempting to decipher the communicative message.
2. Contextual information might not have been sufficient to allow for success in guessing.
3. Guesses might have been wrong altogether.
4. The sheer quantity of target vocabulary would have demanded an unrealistic massive exposure. According to Zahar, Cobb, & Spada (2001) it would take 29 years for 2000 words.

Therefore, Hulstijn and Laufer (2001), Laufer (2006), and Mondria (2003) argued that a form-focused instruction was indispensable and vocabulary learning required explicit tasks. Exposure to input is not sufficient for L2 acquisition in "any learning context that cannot recreate the input conditions of first-language acquisition" (Laufer, 2006, p. 162).

Vocabulary must be practiced as many educators and researchers claim and repeated rehearsals seem to be the key factor. The following section will therefore address how research and pedagogy have addressed this issue.

2.3 Repetition and spaced learning research

This section presents the theoretical background of repetition and focuses on the aspect of spaced learning. It has been informed by cognitive psychology research findings. Next, it addresses how these relate to SLVA,

followed by a brief presentation how this has transferred into some L2 learning systems.

2.3.1 Research findings on spaced learning

As shown in many research findings repetition is an important factor in language learning. Many learning strategies promote repetition and therefore recommend the use of flash cards, vocabulary lists, repeated exposure to L1-L2 word pair translations, self-testing procedures, or pronunciation drills.

It is also agreed that *massed repetition* will not lead to better results. For example, the repetition of a vocabulary item and its corresponding L2 representation twenty times in a row will not lead to a twenty times higher retention rate (Balota et al., 2007; Cull, 2000). *Spaced learning*, however, also referred to as *cyclical learning* or *distributed practice*, leads to higher retention (Baddeley, 1990; Bloom & Shuell, 1981; Carpenter & DeLosch, 2005; Cull, 2000; Dempster, 1987; Nation, 2001), especially when paired with the concept of *desired difficulty* or the *testing-effect* (Bjork, 1994; Karpicke & Roediger, 2007a, 2007b; Roediger & Karpicke, 2006).

The basic idea of *spaced learning* is to return to a previously learned task after some time has passed. Its purpose is robust retention in the long-term memory. How this is done, however, may vary in its purpose and in its realization:

1. The same item is repeated, i.e., L1 to L2 translation of a lexical item. (Ebbinghaus, 1885, 1913).

2. The scope of an item's knowledge is expanded while rehearsed (also referred to as deeper processing). Its purpose is to expand either breadth or depth of knowledge. Information is not only processed repeatedly, but more information is added. For example, a different context for a

vocabulary item is provided, different sample sentences are used in the repetitions, intercultural aspects are added, or grammatical features addressed.

3. The presentation of the rehearsal task may vary: testing, delayed encounter, or simultaneous encounter. The delayed encounter's purpose and that of testing is to create a learning task environment that leads to a higher level of attention and effort, resulting in a higher retention rate (Baddeley & Wilson, 1994, Bjork, 1994; Carpenter & DeLosch, 2005; Dempster, 1987; Landauer & Bjork, 1978; Nation, 2001; Roediger & Karpicke, 2006).

4. Modes of representation may vary. For example, they could include visual cues paired with written representations (Carpenter & DeLosch, 2005; Rimrott, 2009; Yoshii, 2006).

Research on memory has addressed many of these issues; and for the purpose of this study, findings related to interval lengths are my main focus. Interval lengths may vary in any of the above mentioned conditions. They can be random, equal, or graduated (Carpenter & DeLosch, 2005; Cull, 2000; Landauer & Bjork, 1978; Roediger & Karpicke, 2006). The research findings regarding the optimal interval lengths have been controversial and not conclusive. Cull (2000) defined the purpose of expanding recall sessions as the "preventive maintenance of information by maintaining high rates of successful retrieval throughout the test pattern and gradually increasing the difficulty of retrieving item information in order to strengthen storage"(p. 232). This definition corresponds in part with the concept of Pimsleur's (1967) graduated interval recall. According to Pimsleur, an item had to be revisited as long as 60 %

of its knowledge was still accessible. Pimsleur based his hypothesis on Ebbinghaus' (1885) research on the *forgetting curve* described below.

Most of the following research projects have been conducted within cognitive psychology research contexts. However, their findings have considerably informed SLVA. One of the first cognitive psychologists to research retention and the spacing effect and to apply scientific research tools to the concept of memory was Ebbinghaus in 1885.

Though not explicitly set up as SLVA research, Ebbinghaus' findings have often been cited in works on memory and spaced learning (Baddeley, 1990; Bahrick et al. 1993; Karpicke & Roediger, 2007a; Landauer & Bjork, 1978; Schmitt, 2000).

Ebbinghaus used the common fact that a task relearned is done so with greater ease. He set out to research if this had an underlying scientifically measureable pattern. Over a period of 5 years he conducted a series of self observation studies of memory and rehearsal. Having created test corpora of short (CVC-cluster[10]) nonsense syllables, he then memorized sets of 8, 12, or 16 of them and measured the amount of time he needed for these rehearsals. In later experiments, he applied his experimental observations to the memorization of poem stanzas. Interestingly, he found that he could decrease the number of rehearsals from the 6 required for series of nonsense syllables to 4 for stanzas.

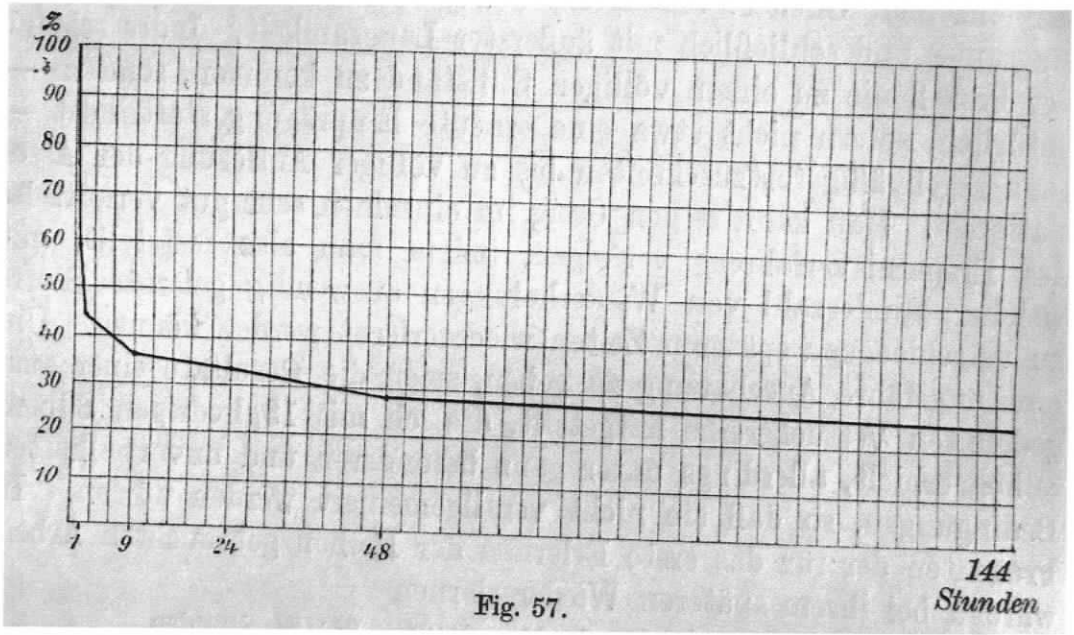
Ebbinghaus then set out to observe the process of forgetting. He measured the seconds he saved when relearning a previously accomplished task. He defined the learning task as completed as soon as he had achieved the objective of two errorless reproductions of the learned item. In various test sessions he addressed the following research question. How much time and learning effort can be saved after the repetition with spaced intervals of

[10] See Ebbinghaus' description of his consonant-vowel-consonant cluster corpora, Ebbinghaus, 1964, p. 22.

20 minutes, 1 hour, 9 hours, 1 day, 2 days, 6 days, and 31 days (Ebbinghaus, 1885, p. 65)? He compared the learning time with the re-learning time and measured the seconds he had saved when relearning the material.

Many years later, in 1913, he published a conversion of this data into a table, commonly known as the *forgetting curve* (Ebbinghaus, 1913, p. 722).

Figure 5: Ebbinghaus' *Forgetting Curve* as shown in Ebbinghaus (1913, p. 722)



Ebbinghaus claimed to have identified a learning curve and a spacing pattern that was relevant to all learners and which could be calculated as a mathematical formula. Even though his research was based on experiments with him as the sole subject, his findings were judged to lead to a general pattern of retention[11]. Ebbinghaus' results have fascinated researchers. They have sparked an immense interest in memory research. Bahrick et al. (1993) reported on more than 300 studies, yet, few studies have contrasted uniform and graduated

[11] Ernest Hilgard (1964), Introduction to Dover Edition, *Memory. A contribution to experimental psychology.* p. VIII.

interval conditions. Landauer and Bjork's (1978) study is among the most prominent addressing this issue.

In 1978, Landauer and Bjork conducted their experiments on optimal rehearsal patterns. They tested massed versus spaced rehearsal, the variants of spaced rehearsal: expanding spacing versus uniform spacing and the effects of practice presentation only versus presentation combined with a test. The participants, 468 psychology students at the State University, NY, were requested to remember 12 target items (names) that were presented in a deck of 50 name cards. The position of these target items within the deck of cards constituted the rehearsal pattern. All cards were presented for 9 seconds each. Therefore, all test series totaled in a rehearsal period of 450 seconds. This was followed by a 30 minute distracter, before the final test was administered. The intervening items for the 3 rehearsals were arranged in different patterns. This then led to a distribution of intervals. The table in Appendix A shows their calculation in detail.

According to Landauer and Bjork's (1978) findings, massed rehearsal was the least effective, uniform spacing was slightly more efficient for presentations with practice only, whereas expanding rehearsal was more effective for rehearsals combining practice and tests.

In 2000, Cull conducted four test series with 66 psychology students at Loyola University in Chicago. He examined the effects of repeated testing and the benefits of testing and cues in spaced learning and therefore set up the target item encounter conditions as: study only—test only—study and test. These conditions were then presented as massed encounters (0-0-0) with no intervals separating the rehearsals, as uniform interval spaced encounters (5-5-5) with five filler intervals between each of the encounters, and as a graduated rehearsal pattern with filler interval spacing set at 1-5-9. The spacing effect for the interval lengths was created by the presentation of filler items, which were 8 seconds

long each. In Cull's first test series, which replicated a learning situation such as the memorization of multiplication tables, his participants were asked to learn a list of 32 word pairs (common English word – uncommon English word[12]). Every target item was first studied for 8 seconds and then represented according to the spaced intervals and the study conditions. (See Table A2 in Appendix A for the calculation of the spaced intervals). Finally, following a one minute distracter, students were tested. Cull's second test series echoed a learning situation comparable to a reading task followed by review questions where students process information and are then tested. This second test setup differed from the first in that it allowed students to study the 8 target items in a distributed practice session at their own pace for 400 seconds prior to the rehearsals. Following a distracter of 1-10 minutes, the 8 items were presented for 12 seconds each. They were spaced at intervals of massed rehearsal (0-0-0), uniform intervals (3-3-3), and graduated intervals (0-3-6). Now, distracter tasks, introduced as word knowledge rating lists, which each took 2-3 minutes to complete, established the interval lengths. (See Table A3 in Appendix A for a calculation of total testing and review time).

Cull's third and fourth studies are particularly of interest to SLVA because these were among the few studies that replicated a "more educationally relevant learning situation" (Cull, 2000, p. 224). They used actual vocabulary words and spaced multiple review sessions across days. Forty-two participants, psychology students at Loyola University of Chicago, were assigned to three spacing conditions: massed (0-0-0)[13], uniform intervals (2-2-2), and graduated practice intervals (1-2-3). Forty vocabulary words paired as common word and uncommon word items were first practiced in a computerized flashcard

---

[12] i.e., the word pair: print– bairn.

[13] Cull had 2 test groups work on a massed schedule. See Table A4 in Appendix A

procedure for 30 minutes. Wrong responses were presented repeatedly to ensure that all target items had been learned. Next, review booklets were distributed. They contained the 40 target items in four study conditions: 10 items study only, 10 items test only, 10 double cards for study–test, and 10 items without review. Furthermore, these booklets arranged the 40 target items according to the conditions of massed, uniform spaced and expanded presentation. Expanding gaps were set at 1 day, 2 days, and 3 days between the encounters of the learned material (denoted 1-2-3). The uniform gaps were spaced at 2 days each (denoted 2-2-2). Three days after the end of the rehearsal period[14] the final cued recall test was administered (See Table A4 in Appendix A for a more details).

Cull concluded that spaced encounters were more efficient than massed encounters. He also found that subjects performed equally well with uniform spacing and with graduated spacing. But, he argued for further research because the repeated presentation of the learning material in the expanding condition might not have been close enough to the tests for students to benefit from this condition. He therefore suggested further studies with the first tests spaced minutes or hours from the initial study period (Cull, 2000, p. 232).

In 2005, Carpenter and DeLosch explored spacing effects in name learning. Their test groups were 62 undergraduate students at Colorado State University. In their experiments they presented 30 name–face pairs in a sequence of 6 seconds each. Filler items, each 6 seconds long, achieved the desired spacing schedules: massed (0-0-0), uniform (3-3-3), and graduated (1-3-5 and 3-5-7). Carpenter and DeLosch' (2005) study could replicate Cull's (2000) research and extend the spacing research findings to non-semantic stimuli processing. They

[14] The most important difference between experiment 3 and 4 is this period. In experiment 4 the final cued recall test was administered 8 days after the rehearsal period.

concluded that spaced items were retained better than items practiced on a massed schedule. However, unlike Landauer and Bjork's (1978) findings, graduated conditions did not lead to higher retention rates. This applied to both expanded conditions. Nevertheless, they advocated further research to determine "boundary conditions" (Carpenter & DeLosch, 2005, p. 633) that would then elicit a positive graduated retrieval effect.

In 2007, Karpicke and Roediger, too, investigated if expanding retrieval practices promoted retention better than equally spaced practices and conducted their research at Washington University with 48 undergraduate students. Their experiments used a series of 52 vocabulary word pairs, 36 tested and 16 fillers, set in the 3 practice conditions: massed (0-0-0), expanded (1-5-9), and uniform (5-5-5)[15] (see Table A5 in Appendix A). Each word pair study or test trial lasted 8 seconds. The first encounter consisted of a study period; subsequent encounters were designed as tests, both with and without follow-up feedback. The presentation of these study and test trials was arranged as a sequence of 121 items with each condition occurring six times. Then, after a 10 minute distracter, respectively a 2-day interval, students were asked to perform a final test.

Karpicke and Roediger (2007b) concluded that an equally spaced retrieval practice produced a higher learning gain on tests performed after a 2-day delay, whereas the graduated schedule resulted in short-term benefits in tests administered with a 10-minute delay. Yet, they argued that any benefit was not a matter of spacing, but rather due to the delay of the first test trial. To back this claim, Karpicke and Roediger measured their students' response latencies. According to their theory, participants' slower response time was an indicator for a deeper processing of the material. This processing would then result in a higher

[15] They also set up two control conditions: one as a study followed immediately by the test and another as a study followed by a delayed test.

retention rate. They argued that contextual elements might have had more time to develop between study period and test. This concurs with notions of *depth of processing* (Craig & Lockhart, 1972), *the desirable difficulty* (Bjork, 1994), and the concept that *attention* and effort put into a task will lead to higher retention (Schmidt, 1994). Karpicke and Roediger's third experiment series explored this hypothesis. It differed from their first two series in that it provided a delayed first test condition. These intervals were scheduled as (0-1-5-9); (0-5-5-5), (5-1-5-9); (5-5-5-5) (see Table A6 in Appendix A). Karpicke and Roediger (2007b) concluded that the crucial factor was therefore not the spacing schedule itself but the timing of the first test and argued: "Delaying the first test improved long-term retention, regardless of how the repeated tests were spaced (p. 704)".

Spaced learning research summary and discussion

The above presented research is now discussed from three angles: long term retention, interval length variants, and educational relevance to SLVA.

First, few researchers were able to test their subjects' knowledge of long term retention like Bahrick et al. (1993) did. Most follow-up recall tests were administered within a time frame of 1 minute (Dempster, 1987), 5 minutes (Carpenter & DeLosch, 2005), 30 minutes (Landauer & Bjork, 1978), 10 minutes and 48 hours (Karpicke & Roediger, 2007a, 2007b), and 3 minutes, 1 hour, 3 days, and 8 days (Cull, 2000). The total time of rehearsal sessions is often measured in seconds.

Second, researchers generally agree that spaced learning is more effective than massed learning. But, the benefit of graduated interval recall over uniform spacing is still a matter of controversy (Carpenter & DeLosch, 2005; Cull, 2000; Landauer & Bjork, 1978) and, comparing these conditions has not been the focus of many studies (Karpicke & Roediger, 2007b). Often, other conditions, for example the testing effect, are examined at the same time. This leads to small

test corpora and small group sizes for the test condition uniform spaced and graduated spaced intervals (i.e., Karpicke & Roediger–4 students per condition and sets of 6 (7) target items).

Most importantly though, there seems to be no reference at how long these graduated intervals should be. This omission seems to be one important condition that is missing for the objective of robust learning and time efficient learning in SLVA. What constitutes the interval length for the threshold level of forgetting (60 % still in the memory), which enables us to retrieve our knowledge of once practiced target items without the time loss of *overlearning*? And, is this figure of 60 % still valid today? The interval spaces as presented in the above discussed research differ considerably. Must we therefore assume that they do not matter as long as they are spaced? Looking back at the neurophysiological processes involved in engraining memories in our minds, we have seen three different modes at work: short-term memory, working memory, and the most time consuming process: long-term memory storage. Neurobiological processes range from electrochemical neural activity to restructuring a complex design of synapses. This again would indicate that most of the experiments discussed above fall into the range of working memory retention. It therefore may not come as a surprise that in Cull's first experiment (Cull, 2000, p. 218) with a total processing time of interval lengths and encounters of mere 144 seconds for both interval conditions he was not able to distinguish a significant difference between uniform intervals and expanded intervals. Their setup was too similar. The same applies for his second experiment. The calculation of total encounters revealed a difference between expanded rehearsals und uniform rehearsals in the second phase only. Again, these are mere seconds compared to the overall one hour length of this test setting.

Cull's third and fourth experiments, both with much longer intervals and with a delayed test, seem to support his claim that there is no difference. Unlike Cull (2000) and Carpenter and DeLosch (2005), who argued this might be due to the fact "that the difficulty of successive retrieval attempts might not have been greater for the expanded (1-3-5) interval than the uniform interval (3-3-3)" (p. 632), I argue that in this scenario they might as well have been both too difficult because they both failed to present the learning material within the sensitive time frame of optimal retrieval. Therefore, their results do not differ. The question arises: Would we reach different results comparing uniform spacing and graduated spacing when we adjust their spacing towards intervals that correspond more to the *forgetting curve* of Ebbinghaus?

Furthermore, a comparison of interval lengths across various research studies revealed that the time lapse between last rehearsal session and final testing – filled with various distracters – is in itself a graduated recall feature. This makes it even more difficult to provide a clear-cut distinction between the two research settings of what determines a uniform spaced and what constitutes a graduated spacing.

Finally, one major difference between the studies above is striking. Because they measure forgetting, not learning, the learning phase preceding the experimental phase is arbitrarily set. Precondition of the research phase is often defined as a learning task completed or as an errorless recall. However, the learning phase tasks in the studies above vary considerably in scope, mode of presentation, size, and time constraints: i.e., Cull's (2000) second experiment with 400 seconds for 38 items, a self-paced setting, and unlimited access; Cull's (2000) third experiment with 30 minutes, unlimited attempts for 40 items in random order and tested; Bahrick et al.'s (1993) study with unlimited time, self-paced, and unlimited attempts for sets of 50 items; Landauer and Bjork's (1978) research

with 450 seconds for 50 items, limited attempts and a controlled presentation. Findings therefore do not compare easily.

The learning phase is considered to be a reference point from which further rehearsal patterns are then tested. However, would it not make a difference what, how and for how long a subject practiced? I argue that the learning phase should not be seen as a preceding procedure only. It is already part of the experiment itself.

Research findings on the concept of *attention* and the *depth of processing hypothesis* point into this direction (Schmidt, 1994; Schmitt & Schmitt, 1995). Schmitt and Schmitt (1995) argue that "the depth of processing hypothesis states that mental activities which require more elaborate thought, manipulation, or processing of a new word will help in the learning of that word" (p. 135).

Furthermore, the kind of target items we choose to retain may have an effect on the learning gain. We know that certain parts of speech are easier to learn than others.[16] The choice of the test corpora must reflect this. Ebbinghaus may have chosen well for his research to use nonsense syllables. But nonsense syllables do not have a semantic reference point as L2 vocabulary items do. This semantic reference in L1 aids comprehension. L1 and L2 acquisition differ because we construct a link between what we know in L1 with what we intend to learn in L2. This leads to the question: Will this semantic reference point speed up retention because my knowledge of L1 helps me to retain an L2 vocabulary item?

In addition, the interval frequency we encounter in these studies varies considerably. Yet, we know very little about optimal numbers. Rehearsal sessions are often chosen in sets of three (Dempster, 1987; Cull, 2000; Landauer & Bjork,

[16] See Nation's (1990) and Singleton's (1999) evaluation of the ease learning cognates; Ellis and Beaton's (1993) study on differences between the word categories nouns and verbs; Aitchison (2003) on the distinction between content words and function words; and Schmitt's (2008) research review of studies on the aspect of word-form with the regard to the ease of retention.

1978). Bahrick et al. (1993), however, practice 13 times in their first test series and 26 times in their second experiment.

In summary, the spacing effect has proven to be a well researched phenomenon. According to Bahrick et al. (1993) this research topic has been investigated by more than 300 studies in cognitive psychology. This impressive number has undoubtedly risen in the meantime. These studies in the context of cognitive psychology research were designed for carefully constructed and controlled, yet artificial learning environments. Not only is motivation extrinsic (i.e., the participation in the test allows for course credits) and not intrinsic as – hopefully– expected in a language class, the test corpora are also arbitrary: names, nonsense words or non-frequent words of English. These do not compare easily with the learning conditions of SLVA because using them strips the testing condition of one important L2 learning condition: the ability to connect what is already known with what we intend to learn. Using cognitive psychology research to come to an understanding of SLVA processes may guide SLVA researchers but does not suffice to explore L2 learning. According to Bahrick et al. (1993) and Cull (2000) the graduated spacing effect researched in longitudinal studies of SLVA has not been the focus of many studies.

The following section presents how pedagogy has been informed by this research.

2.3.2 Spaced learning applications in SLVA

Spaced learning research in cognitive psychology has received much attention in education where it has sparked considerable interest in applying its findings in educational settings. Yet, education has continued to base the development of much of its L2 learning material on those findings that were

obtained decades ago and relied on Ebbinghaus' (1885) work (i.e., L2 computer learning software "az6-1"; L2 computer software "Phase 6"; Leitner's "Lernkartei"; computer software "Inquisitor"). Other L2 learning material development was informed by Seibert (1927) such as "Pimsleur's Approach" (1967). I argue that it is worthwhile to look into research findings, which were obtained more than a century ago, and to evaluate if they still apply to today's learners and today's learning environment.

Nation (2005) promoted direct learning of vocabulary with word cards as an efficient learning device. He recommended that this method of direct learning and repetition should be part of an overall vocabulary learning agenda. Hulstijn and Laufer (2001) pointed out that an explicit memorization stage following other strategies such as inferring or verifying would greatly improve retention. Their view is shared by many others (Huckin & Coady, 1999; Laufer, 2006; Mondria, 2003; Mondria & Mondria-deVries, 1994; Scherfer, 1994; Schmitt, 2000, 2008).

Oxford (1990) promoted a staggered vocabulary practice in her renowned SLA textbook. She suggested practicing 7 times on a graduated interval schedule (15 min; 1 hour; 2 hours; 1 day; 4 days; 1 week; 2 weeks).

The following provides a brief presentation of some SLV approaches and learning systems that are widely used today and that are based on spaced learning.

Word cards and memorization systems revisited

The following examples of learning systems have been chosen because they follow the principle of spaced learning and are relevant to the research questions of this study. Word cards, vocabulary flash cards, and variations of memorization systems have been widely used in the past (Schmitt, 2000). It is worthwhile to explore if they indeed accomplish what they set out to do.

Pimsleur's graduated interval recall – 1967

Pimsleur (1967) designed an audio-lingual language learning system following four learning principles[17]. He named the concept of *graduated interval recall* as his second principle.

Nation (2001) pointed out that Pimsleur based this principle on his own observations of his Greek language classes. He did not conduct experiments of his own but referred to research of memory and retention of Ebbinghaus (1885) and Seibert (1927). The purpose of Pimsleur's spaced repetition was to move vocabulary items into the long term memory. Pimsleur defined how these rehearsal sessions should be scheduled very precisely. He based these 11 intervals on an exponential of 5. (5^2, 5^3, 5^4, ...) leading to the schedule below.

Table 1: Pimsleur's practice schedule

frequency	1	2	3	4	5	6	7	8	9	10	11
interval lengths	5 s	25 s	2 min	10 min	1 hr	5 hrs	1 day	5 days	25 days	4 months	2 years

However, we do not really know why he chose the exponential of 5 and why he set up 11 practices. A closer look at his teaching materials even revealed that the intervals were difficult to time in the way his approach required.

Furthermore, he stated that knowledge of a word meaning could be retrieved to 100 % if the item was revisited as long as a minimum 60 % of its knowledge was still remembered. He argued that after each repetition the decrease of knowledge was less rapid and therefore the time lapse between each revision session could be increased, hence leading to the graduated interval.

---

[17] 1. Anticipation

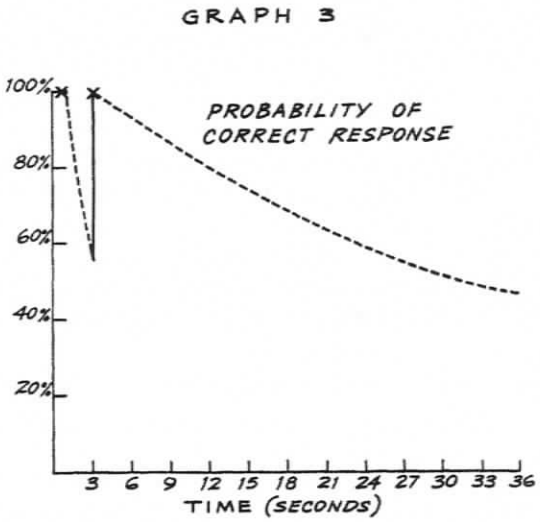
2. Graduated interval recall

3. Core vocabulary of high frequency words

4. Focus on auditory skills.

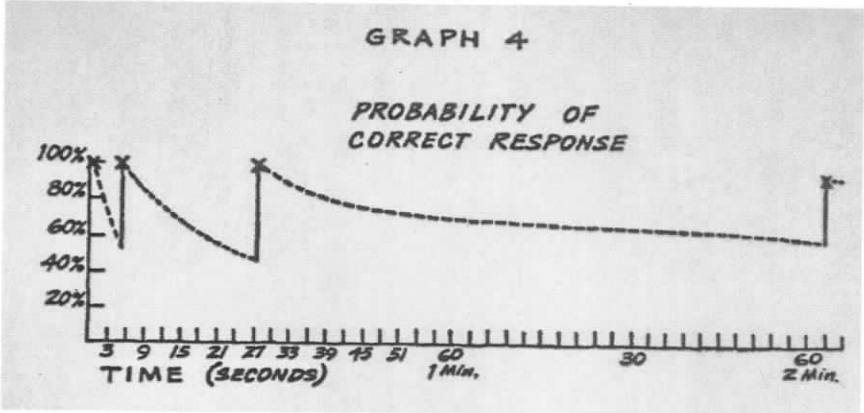
But, Pimsleur did not report why this minimum retention value was set at 60 %. Even though he admitted that some words such as cognates might not need as many repetitions as long, infrequent non-cognates, he assumed that there was an ideal schedule adaptable and applicable to all learners and all lexical items.

Figure 6: Pimsleur's retention curve as shown in Pimsleur (1967, p. 75)



Based on this assumption he calculated the following graph of correct response probability.

Figure 7: Probability of correct response as shown in Pimsleur (1967, p. 75)



Language programs are now sold – more than 30 years after his death – as "Pimsleur Approach Learning System" for many languages. They are based on 30-minute audio sessions of structured repetitions.

Leitner's hand computer "die Lernkartei" –1972

Figure 8: Leitner's hand computer "*Lernkartei*"

Another prominent learning device based on Pimsleur's and Ebbingshaus' concept of structured cyclical repetition was Leitner's *hand computer*. In 1972, Leitner created the hand computer or as Leitner himself calls it in German "die Lernkartei", a memorization device consisting of flashcards and a box with 5-6 sections of progressively larger size. The term to be learned is written on one side of the flashcard, the prompt on the other. Ideally, the daily input is determined by the corpora size of a daily vocabulary learning routine and should be limited to 12-15 new items. This amount averages the estimated limit for new intake and is usually recommended by the instructor. These flashcards then begin their learning cycle in the first compartment. Users take them out and memorize them. As soon as the first compartment is filled with flashcards these are then reviewed. Correct responses move up to the next compartment. Incorrect ones drop back to the first compartment. The size of these compartments is progressively larger. Thereby, because it takes longer to fill the compartments

and new input is limited to 12-15 items, the review periods are spaced. This then is intended to lead to the same functionality as Pimsleur's (1967) model of *graduated interval recall*.

The flash card boxes have been in use in the German school system for many years. Their usefulness is promoted in teacher training programs (Schröder & Roedig, n.d.). English as a Second Language textbooks publish sets of flashcards to match their corpora or create multimedia vocabulary learning environments based on this type of cyclical learning. Many of these flashcard systems are available both in print and as multimedia devices (i.e., *Pons-Lernbox Englisch im Handumdrehen, CD-ROM Langenscheidt Vokabeltrainer,* and *Cornelsen G2000 Handytrainer "Mobile English"*). Mondria and Mondria-deVries (1994) inferred that these devices must have been a success because they were so widely published. But, apart from concerns about the impracticability of flash cards and issues of how to edit and proof-read content (Lüders, 2005; Mondria & Mondria-deVries, 1994; Schmitt, 2000), there has been no published research on how these flash cards are processed and handled. How can we ensure and prove that the necessary time frames for an optimal recall as postulated in the research findings of memory and vocabulary retention (Ebbinghaus, 1885; Landauer & Bjork, 1978; Leitner, 1972; Mondria, 2000; Pimsleur, 1967; Rott, 1999) are even maintained in the review sessions? Compared with Pimsleur's (1967) precisely defined intervals there are striking differences to Leitner's (1972) flash card system.

Leitner's intervals are crudely scheduled:

1. When students wish to practice and memorize is at their discretion. Even the practice of the first compartment may therefore vary from seconds to days.

2. The amount of words entered will determine the interval: the more entered – the sooner the compartments will be filled and subject to review. This leads to interval variations.

3. Time delays, interruptions, variations in learner effort are not accounted for. This may lead to more items returned to the first compartment, reentering the learning cycle at the beginning because their threshold level of retention was not observed. This will lead to a flooding of the first compartment and again to uncontrolled interval spacing. Mondria and Mondria-deVries (1994) set to address these issues in the development of a computer assisted learning product.

Mondria's CALL program

Mondria and Mondria-deVries (1994) developed a CALL program that was based on a textbook and incorporated the learning system of Leitner (1972). They advocated its use as a less time-consuming activity compared to the flash card writing and promoted its additional features of sound files and image files. Yet again, we cannot know if students use it and how they use it. This in particular, is a research aspect that needs to be taken into account.

For the research presented in this thesis it was therefore necessary to design and program the web-based application software (Vivo) that would not only present and practice the lexical corpora, but would also track and document every user's learning process for every single vocabulary item. Its features are presented in more detail in section 3.4.4.

2.4 Conclusion and Research Questions

This section introduced the SLVA research pertaining to spacing effects and frequency of encounters. First, it addressed relevant interdisciplinary perspectives of memory and retention. Next, it focused on research findings on

spaced learning and rehearsal frequency. Then, it presented various memorization systems and pedagogical implications based on these findings.

It established that most of the relevant research on spaced learning has been informed by cognitive psychology research. Even though the spacing effect was a well-researched phenomenon in psychology, research of specific applications in SLVA has lagged behind.

Most researchers agreed that spaced learning is more effective than massed learning. However, the benefit of uniform spacing versus graduated spacing of practice encounters was still discussed controversially. Whereas Landauer and Bjork (1978), Oxford (1990) and Pimsleur (1967) favoured a graduated schedule, Carpenter and De Losch (2005), Cull (2000), and Karpicke and Roediger (2007b) did not report an advantage of the graduated schedule. This was in contradiction to pedagogical practice promoted in teacher training and SLA textbooks (i.e., Leitner, 1972; Mondria & Mondria-deVries, 1994; Oxford, 1990; Pimsleur, 1967; Schröder & Roedig, n.d.). This contradiction called for further exploration.

Next, the frequency of exposures examined in research varied considerably. Most studies in cognitive psychology research had 3 rehearsal sets scheduled (Cull, 2000; Karpicke & Roediger, 2007b; Landauer & Bjork, 1978). Much of the spaced learning material in SLVA, however, was based on 5 encounters as recommended by Leitner (1972). However, these frequency values were not consistent and other researchers have recommended even more encounters (i.e., Oxford (1990) promoted 7 practices, Pimsleur (1967) suggested 11 encounters, and Bahrick et al. (1993) used 13 and 26 practice sessions in their research). In conclusion, the frequency issue called for further research.

The research presented in this study was therefore designed to answer the following two research questions:

1. Which students would perform better on their tests: students learning vocabulary on a practice schedule with graduated intervals or students who had practiced this vocabulary on a practice schedule with uniform intervals?
2. Would students practicing vocabulary with less practice sessions (2/3 encounters, as suggested by cognitive research) perform equally well or less well than students studying more often (4/5 encounters as suggested in many SLV programs)?

3. Methodology

This present study was carried out in the fall semester of 2008 at the University of Victoria (UVic), Canada. It follows the principles of an experimental design as described by Seliger and Shohamy (1989).

The following section presents the research schedule, the setting, the design, and the data collection procedure.

3.1 Research design and schedule

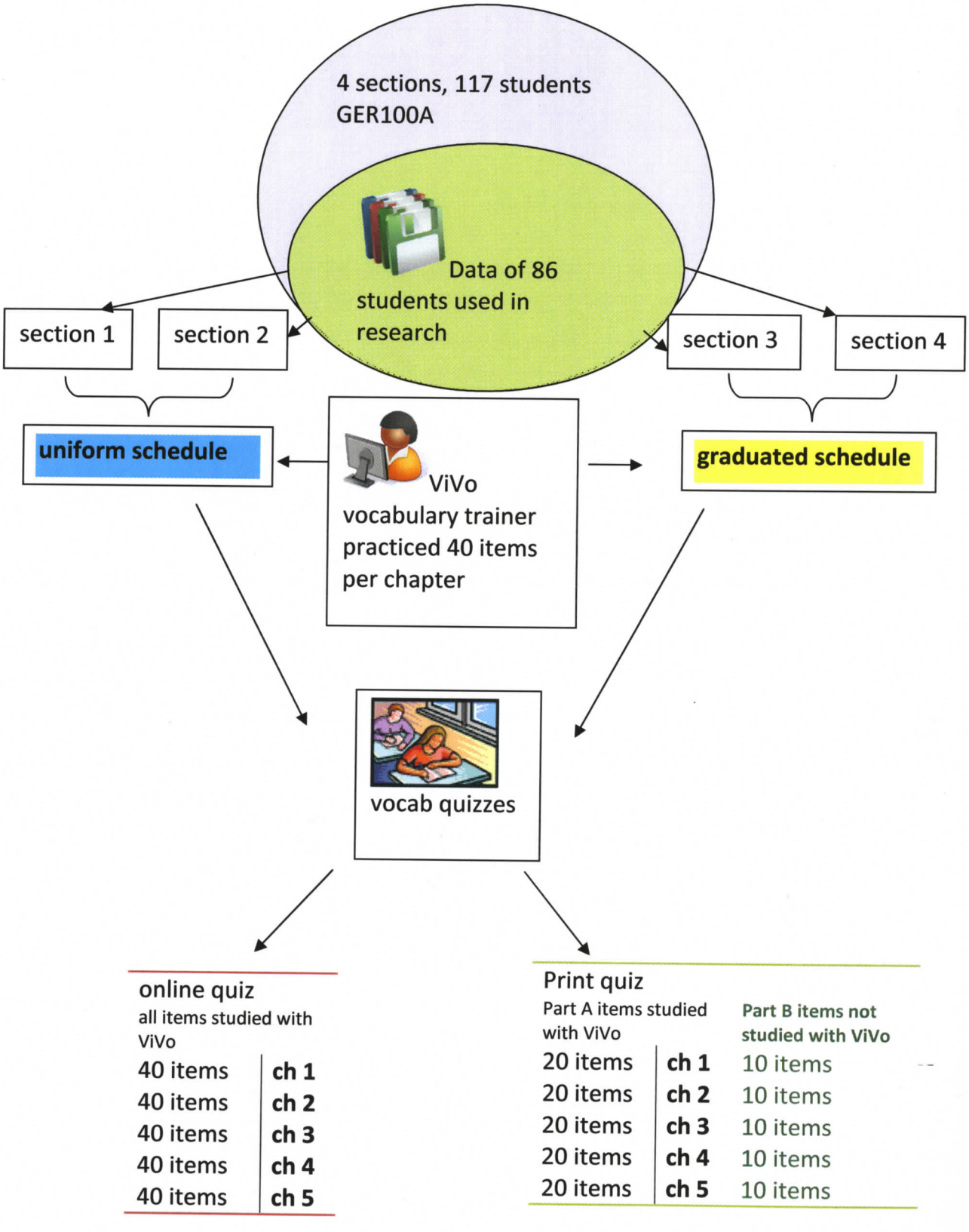
To answer the research questions, I set up a quantitative and qualitative research scenario within a learning environment and used the following research procedures, described in section 3.4 to collect data

- Questionnaires A and B
- the web application ViVo with two different settings of learning cycles
- ViVo user logs
- five vocabulary quizzes in print
- five vocabulary quizzes online

First, the effectiveness of the web application vocabulary learning software ViVo was tested in a pilot study in spring 2008. Then, in the winter of 2008, ViVo was reprogrammed to provide controlled access according to two different test conditions: graduated intervals and uniform intervals. These represented the independent variables.

The graph below provides an overview of the research procedure and setup.

Figure 9: Research procedure and setup



The data of 117 students from four sections GER100A sections was examined in this study. Because students had to comply with filter criteria (see 3.4.2 and 3.4.3) only the data of 86 students was entered into the data analyses. These students were divided into two groups according to two test conditions. One group used ViVo to learn vocabulary items following a graduated spaced learning schedule (henceforth called graduated group – GG). The second group used ViVo to learn vocabulary items following a uniform spaced learning schedule (henceforth referred to as uniform group – UG).

At the beginning of the term the students of all sections filled out Questionnaire A. It provided information on their language background.[18] Then, both groups used ViVo to practice a total of 200 vocabulary items during the winter term 2008. ViVo prepared them for vocabulary chapter quizzes that were graded as part of the students' evaluations.

The vocabulary retention based on the test scores of both groups was investigated. The results of five chapter vocabulary quizzes in print and five chapter quizzes online were compiled. I analyzed these results and correlated them to the participants' ViVo user logs. The students' performance on these tests constituted the research's dependent variable measure.

At the end of the term both groups filled out a second questionnaire, Questionnaire B.[19] It provided information on students' learning strategies and use of online tools. This questionnaire is described in more detail in 3.4.3.

3.2 Setting

This section establishes that both test groups shared the same learning material and background. It describes the course setting, the course objectives,

[18] See Appendix D.

[19] See Appendix E.

and the textbook. It then presents a detailed description of the learner corpora and test corpora that were part of this study.

3.2.1 The GER100A course

The Germanic and Slavic Department at the University of Victoria (UVic) offers a language program for German as a Second Language. Learners with no prior knowledge of German start with a one semester course at the beginners' level GER100A. They then have the option to continue their language program at the low intermediate level of GER100B, the intermediate level GER 200 followed by GER 300 and GER 400. The University Calendar and the department's website publish course outlines with detailed descriptions of course content and objectives.

Usually, sections of GER100A have 30-35 students each. Instruction on basic grammar, vocabulary, fundamental structures of everyday communication, and the presentation of intercultural aspects of German speaking countries are offered to students with no previous knowledge of German. Instructors meet with students three times a week for 50 minutes per class session. Class time instruction takes place in a classroom setting or occasionally in the Computer Assisted Language Lab (CALL facility). Students are expected to work individually in the language lab as part of their course requirement. There, they study for approximately one hour per week towards a total of at least 500 minutes per term. Furthermore, at the beginning of the term they were advised that, in addition to the above, they had to study outside class on a regular basis. Instructors address issues regarding study skills in L2 acquisition at the beginning of the term and revisit these throughout the course.

The students' grades are based on attendance (class and CALL) 10%, in-class participation 10%, chapter online exercises 7.5%, five online vocabulary quizzes

2.5%, three assignments 10%, five vocabulary quizzes 10%, two accumulative chapter tests 20%, and a final exam 30%.

3.2.2 The Textbook Deutsch NaKlar

GER100A courses use the fifth edition of the textbook *Deutsch NaKlar*. The following presents how the textbook implements vocabulary learning pedagogy. In particular, it addresses how vocabulary is embedded in context. It then discusses corpora size and frequency issues within the textbook.

Both test groups used the same textbook: the fifth edition of *Deutsch NaKlar!* The course GER100A usually completes the first five chapters. Its authors Di Donato, Clyde, and Vansant (2007) claim to provide an approach that embraces the five C's of standards for foreign language teaching: Communication, Connections, Culture, Comparisons, and Communities. According to them these permeate the activities, exercises, readings, and language tips in each chapter (Di Donato et al. p. XVII). Some of its claimed hallmarks are the extensive use of authentic material and abundant communicative activities.

The textbook presents vocabulary in a functional framework and introduces it in context: visuals, dialogues, and narratives. At first, this context introduces vocabulary receptively. Vocabulary acquisition then progresses from controlled and form-focused to open-ended and interactive, then it concentrates on production.

Deutsch NaKlar uses various approaches to highlight vocabulary in the chapter text. First, new, active vocabulary items appearing in context for the first time are presented in bold print. Students are expected to infer their meaning from context. Second, language boxes next to new topics aid students to verify

meaning.[20] Third, some, but not all new, passive vocabulary items are italicized. Finally, a textbox called "Sprachtipps" provides further lexicogrammatical information.

The textbook provides dictionaries in various forms. At the end of every chapter a word list compiles the newly introduced active vocabulary. The list is organized in conceptual groups or according to grammatical features. Furthermore, the appendix provides a German-English dictionary of the textbook corpora.

The textbook does not introduce the International Phonetic Alphabet or any other pronunciation guide.

3.2.3 Learner Corpora

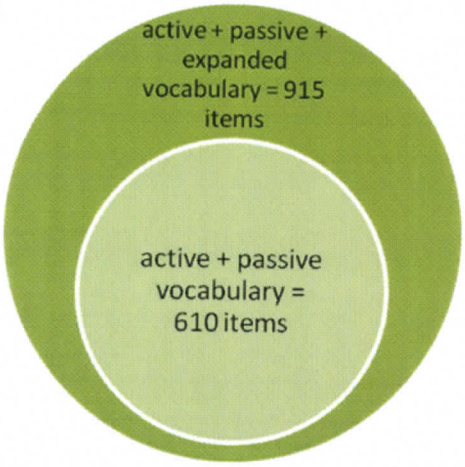
By the end of GER100A, students acquire an active—passive[21] learner corpus of 610 items based on the vocabulary items of the textbook. Additional vocabulary exposure occurs in the authentic material (i.e., ads, menus, and flyers) and in enriched reading texts. The students infer meaning of these vocabulary items by context or look them up in the appendix: the German-English Vocabulary section of the textbook. The ratio of this expanded vocabulary to the active—passive vocabulary measures 1.5: 1.

---

[20] However, it does not explain why students would not simply go to the adjunct textbox to look up the new words instead of inferring meaning from context.

[21] Di Donato et al. (2007) use both set of terms "active—passive" and "receptively—productively" interchangeably.

Figure 10: Ratio of textbook's active/passive vocabulary



These figures seem to imply that students have processed all vocabulary thoroughly. However, a frequency analysis of the textbook's lexical corpora based on the German frequency dictionary by Jones and Tschirner (2006) revealed that incidental exposure to vocabulary items was randomly distributed.[22] For example, chapter 2 introduced the two vocabulary items "nur – only" and "das Zimmer – the room". The item "nur" had a frequency ranking according to the German frequency dictionary of 44, yet appeared only 8 times in this chapter. The item "das Zimmer" had a frequency ranking of 609, but appeared 101 times in chapter 2. How incidental encounters and frequency related to the outcome of the research question is discussed in sections 4.1.1 and 5.2.2.

3.2.4 Test corpora

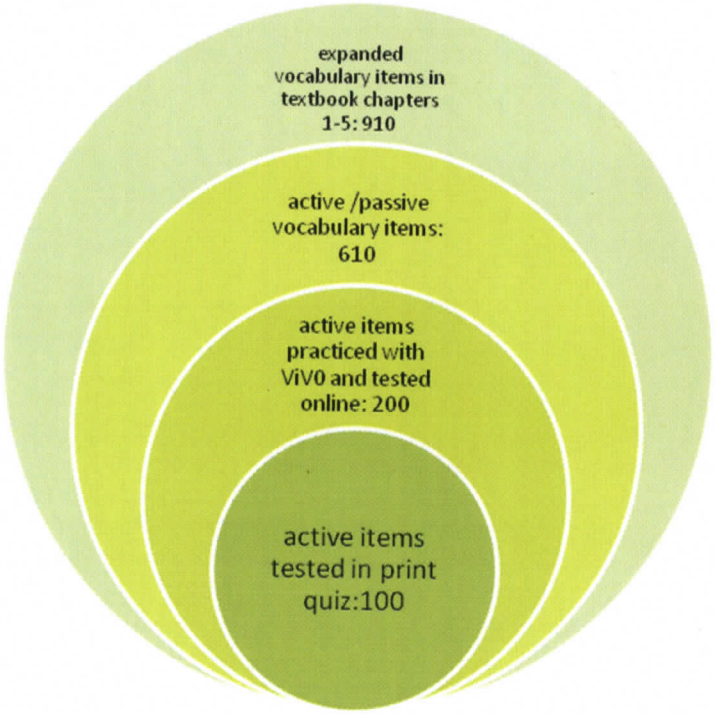
The following addresses the size of the test corpora and the ranking objectives that led to the choice of vocabulary items.

[22] I evaluated the vocabulary quiz test corpora based on their frequency in German and their incidental encounter in the textbook *DeutschNaklar*. See Appendix B and Appendix C for the word lists.

Test Corpora Size

The test corpora of the online quiz and the print quiz were made up of 200 vocabulary items. These were part of the textbook active vocabulary as described above. Participants practiced them using ViVo. All 200 were tested online. 100 of these constituted the items that were chosen for the production part of the print quiz. See Figure 11 below.

Figure 11: Ratio of textbook vocabulary and test corpora



Hierarchy of test corpora choice

This research based the test corpora choice on the following ranked objectives as described further on in more detail. These criteria were set in the following hierarchical order. The vocabulary items, which were chosen,

1. were part of the active learner corpora of the textbook
2. consisted of base words
3. included content words and function words

4. included cognates
5. had a high frequency rate in the German language corpora.

Sometimes this hierarchy was disregarded, when other factors were identified as more important. For example, the cognate "Party" has a frequency ranking of 3634 in the German frequency dictionary (Jones & Tschirner, 2006). In comparison, the word "geben" has a frequency ranking of 57. Even though the frequency ranking of "Party" exceeds what Nation (1990) deems basic vocabulary for speech, it is very likely a word of importance to young adults. This and the fact that it is a cognate take precedent over the frequency within the German language corpora.

The paragraph below describes these objectives for the test corpora choice in more detail.

Textbook's active learner corpora

I chose a total of 40 vocabulary items from each chapter of the textbook *Deutsch NaKlar*. They were part of the chapter's 110–120 active vocabulary items and concurred with the objectives listed above, but only these 40 were submitted to the learning cycle of ViVo. The instructors informed the students that part of their graded print quiz would test 20 of these 40 words. I chose 40 items for two reasons. First, following Baddeley's research on the concept of a *limited attentional capacity* in a stressful situation (Baddeley, 2007, p. 260), this number is too high for a short-term last minute cram prior to the quiz. Therefore, students could not rely on short-term memory performance (Baddeley, 2007; Cowan, 2001, 2005). Second, as a pragmatic consideration, students could practice 40 items within a reasonable time in one practice session.

Of course, students encountered these items incidentally as part of their learning experience in class. However, it was highly unlikely that they would have

had access to these items other than that provided by the learning experience in a German beginner's class. Questionnaire B aimed at identifying additional resources students had accessed so that they could be taken into account when discussing the data (see Appendices E and F).

Base words

Nation (1990) defined the term base word as "any word whose meaning cannot be predicted on the basis of its components". Following this definition, I only chose base words. Furthermore, I did not choose words that were similar and allowed to derive meaning. For example, I did not choose the words "*der Arbeiter* – the worker" nor "*die Arbeit* – the work" because I had used "*arbeiten* – to work".

Content words and function words

The test corpora included function words and content words. Friederici (1985) has described function words as items linked to syntax and content words to lexical properties. Aitchison (2003) refers to content words as the "bricks" and function words as the "mortar". However, as she and others (Singleton, 1999) cautioned, the distinction may not be as clear-cut as previously assumed. Singleton (1999) stated "We reject the implication that there is a hard-and-fast distinction between 'content words' (or 'full words') and 'grammatical words' (or 'empty words') which runs counter to what most lexicologists now seem to be saying on the matter" (p. 255). Consequently, the compilation of the test corpora based on the distinction between content words and function words could only follow a rough categorization of these items.

The chapters covered a variety of topics, but not all chapters lent themselves to an equal amount of available content words versus function words. Nevertheless, I set their ratio at approximately ¼ function words to ¾ content words. The test

corpora choice also roughly reflected this ratio: 50 function words and 150 content words. The word categories nouns, verbs, adverbs, and adjectives provided the content words. The word categories prepositions and pronouns supplied the function words.

Cognates

The total number of items in the test corpora was 200. Of these 200 words 51 were cognates and close-cognates. Carroll (1992) defined cognates as lexical L1 and L2 items that bilingual speakers regard as "being the same thing". She described them as words with formal similarities that share a semantic closeness in most cases. Her definition [23] has informed this study with the exception that I have only used cognates for this research that were semantically identical (i.e., "die Lampe – the lamp"). Close-cognates differ in their presentation either phonologically, or in their written form, but they would still be recognizable as a similar form by L1 speakers. (i.e., shoe –"*Schuh*" sound alike, but they differ in spelling; minute – "*Minute*" are spelled the same but differ in pronunciation). Whenever possible, cognates and close-cognates were selected from all word categories. Even though acquisition differences between function words and content words were not part of this research, the discussion in 5.3 addresses these as interrelated acquisition factors. Furthermore, this corpora design prepared for further data evaluation in future research with word category differences as research topic.

---

[23] Cognates are "(...) lexical items from different languages which are identified by bilinguals as somehow being 'the same thing'. Cognates have at least 4 essential properties: 1) they are always structural units; 2) they are words; 3) words paired may be but need not be semantically identical; 4) there is always some kind of formal resemblance between cognates." (Carroll, 1992, p.93).

Frequency in German language corpora

The test corpora were balanced towards more frequent words. I used the German frequency dictionary by Jones and Tschirner (2006) to determine the frequency of the 200 vocabulary items in the test corpora (see Appendices B and C).

This motivated students to learn because these words mattered. According to Nation (1990), 95% of written texts are comprised of 4000-5000 most frequent words and 85% of speech uses 1000 of the most frequent words. Therefore, it made sense to take frequency into account when creating vocabulary corpora for L2 learners. Exceptions were commonly known cognates as discussed above.

3. 3. The participants

This section briefly describes the social and educational background of the participants. It then explains how I encouraged their active participation in this study.

3.3.1 Student background

The participants, whose data was included in the analyses for the research questions, were 86 (48 female/ 38 male) university students in four German beginners' classes[24] within the context of an English speaking environment. They had had no prior German language instruction or knowledge of German. Therefore, their vocabulary retention could be attributed to their instruction within the context of this course.

The subjects of the four sections were divided into two test groups. The group with the testing condition of graduated intervals (GG) had 40 participants.

[24] The four classes had a total of 117 students. Thirty-one did not qualify for participation in the research because they matched one of the filter criteria (see 3.4.2; 3.4.3; 3.4.4).

The group with the testing condition of uniform intervals (UG) had 46 participants.

The students completed a questionnaire on their social and educational background. Therefore, I could provide evidence that the student profiles of both groups' participants resembled each other. The filter criteria to establish the homogeneity of the two test groups are described in more detail in sections 3.4.2 and 3.4.3.

3.3.2 Student participation

I conducted this study under the conditions of a university credit course. This necessitated participating students' consent and active cooperation at all times. Students signed an ethic consent form after they were introduced to the objectives of this research, the features of the online trainer, and relevant research findings on behalf of spaced learning benefits. They were also informed that this vocabulary training would be evaluated as part of their online activity grade.

This study's success depended on students' cooperation and motivation to do their assignments according to their schedule. Therefore, the following measures were taken to encourage their participation. First, students received a printed schedule at the beginning of the term. Next, their instructors reminded them of the encounter dates. Furthermore, their course management system Moodle had these dates in its task calendar, and students received an email reminder prior to their encounters. In addition to the above described measures, participation draw prices were introduced as incentives for task completion.

3.4 Data collection

This research used the software ViVo, two questionnaires, 5 print quizzes, and five online tests for data collection. Furthermore, the textbook *DeutschNaKlar* was scanned for frequency of the test corpora items. This amounts to the following data sets processed in this research:

Table 2: Data collection means

Source of data	Size
Textbook frequency scan	100 print quiz items
Questionnaire A	117 with 11 questions each
Questionnaire B	117 with 16 questions each
ViVo	430 user log entries
online quizzes	233 quizzes with 40 questions each
Print quizzes[a]	361 quizzes with 30[b] questions each

Note. [a] These figures for print quiz and online quiz were calculated based on the filter criteria described in this section.

[b] Twenty questions were based on the test corpus practiced with the online trainer; ten test items were part of the chapter vocabulary list not practiced with the trainer. See 3.4.5 for a detailed description.

Table 3 summarizes the purpose of these data collection methods and names the data or measurement provided by them. These methods are then discussed in more detail in the following sections below.

Table 3: Data collection instruments and their research purposes

	Instrument	Issue addressed	Purpose	Data or measurement provided
1.	Textbook scan	frequency of test corpora	validity: to control incidental encounter variables	frequency list of test corpora
2.	Questionnaire A	filter criteria, internal validity	to establish homogeneous groups	social profile data
3.	Questionnaire B	filter criteria, internal validity, qualitative data	a. to establish homogeneous groups with similar learning strategies b. to elicit qualitative data on students' learning techniques	student responses
4.	ViVo user log	filter criteria, internal validity, research question 1 and 2	to document use of online tool	user log documentation
5.	5 online vocabulary quizzes on Moodle	research question 1and 2	to compare performance of UG and GG	user logs with test results
6.	5 print vocabulary quizzes	internal validity, research question 1 and 2	to compare performance of UG and GG; to compare performance of ViVo-trained and non-ViVo trained items	a. test results of items learned with ViVo 2. test results of items learned without ViVo.

The students also encountered the vocabulary items they learned with ViVo in class. Over a period of 14 days students completed activities that introduced and practiced these items within the chapters. These chance encounters could have influenced the students' performance. Because they constituted a possible reliability issue, I examined them more closely (see below and 4.1.1).

Vocabulary acquisition based on incidental encounters is still a controversial debate. Some researchers claimed that a considerable learning effect is achieved by these incidental encounters. If this were true, frequency within the textbook would bias the results of vocabulary learned with the online tool. I therefore scanned the textbook Deutsch NaKlar and performed a textbook word count of those 100 vocabulary items that were tested on Part A of the print quiz (see Appendix B). Part A tested the items studied with ViVo. The following describes how the frequency count was executed.

Derivatives and compound nouns that were semantically **close** were included in the count (i.e., *"die Woche" – week* and the compound *"das Wochenende" – weekend*). However, if they were not close enough semantically, they were not included in the count (i.e., the verb *"zeigen"* – to show, and the noun *"die Anzeige"* – the ad). I then compiled frequency results in a list (see Appendix B). Based on this frequency, I then chose a random sample of 6 items: 3 with high frequency, 3 with a low frequency.

Table 4: Textbook frequency sample

German word	das Buch	kaufen	der Zucker	wie	die Wohnung	mit
frequency in chapter	7	8	2	53	41	28

Next, I tracked their scores on the print quizzes and their user entries on the online quizzes. ViVo user logs and Moodle user logs documented user entries for every single vocabulary item. This detailed data therefore enabled me to evaluate incidental learning influences.

3.4.2 Questionnaire A

The purpose of Questionnaire A was to elicit student background information (see Appendix D). I used this information to create social profiles in order to establish homogeneous groups. This questionnaire is a standard procedure in all beginner German classes at UVic. It identifies previous knowledge of German, previous instruction in German, German heritage, proficiency in other languages and provides social data on faculty, major, and years of study. Filter criteria screened student responses. Students who met any of the following conditions were removed from the data pool

1. previous German course at university or college level
2. more than 6 months in a German speaking country after the age of 11
3. three years of German instruction at high school level
4. a combination of the factors:
    a. one parent is a German speaker and 1 or 2 years of high school classes in German,
    b. or more than two grandparents are German speakers and 2 years of high school classes in German.

The four German sections had a total of 117 students. Seventeen of them were removed from the data set as a result of this screening process.

3.4.3 Questionnaire B

Questionnaire B was distributed in the last week of the course (see Appendix E). Its purpose was twofold. First, it established filter criteria so that the students' learning strategies would not interfere with the validity of the research findings. In particular, it asked if students had employed learning techniques that corresponded closely to the design and methodological concept of the online tool ViVo. Second, it provided qualitative data to evaluate the students' use and acceptance of the online tool ViVo.

Three conditions elicited by this questionnaire were used as filter criteria and led to an exclusion of students' data.

1. Students had used another online vocabulary trainer or software (Question 3).
2. They had made extensive use of flash card practice (Question 16). If they had chosen answer 4 as their favourite learning strategy ("I write flashcards and carry them with me to study"), their data was excluded.
3. They had studied the ViVo words on a word list every day (Question 7).

Following the evaluation of this questionnaire, another six of the 117 students were excluded from the data analyses.

Questionnaire B's second purpose was to compile qualitative data. Students' open-ended responses and multiple-choice answers provided information on students' acceptance of the program, how they perceived its user friendliness, and how they rated on-line activities in general.

To this means, Question 3 addressed preference of online tools versus conventional learning methods. Question 5 required students to check those electronic resources they felt comfortable using. Finally, Questions 14 and 15 asked for favourite parts about ViVo and suggestions for improvement. This

qualitative data therefore provided valuable input on students' motivation and attitude towards learning with an online vocabulary tool. These student comments also informed the discussion on behalf of ViVo's user friendliness in section 5.1.

Questions 11, 12, and 13 were used to elicit students' opinions on the number of practice sessions they felt they had needed. This information complemented the statistical analyses of the second research question as discussed in section 5.3.

3.4.4 Vocabulary learning web application and research tool ViVo

Implementing a software tool such as the vocabulary trainer ViVo as means of data collection necessitates proof that this tool will perform as intended – namely, help students study vocabulary efficiently. Many research studies have established the benefit of a CALL environment (i.e., Baturay, Yıldırım & Daloğlu, 2008; Jones, 2006; Groot, 2000). The following text presents how a pilot study verified the online tool ViVo's efficiency. It then describes features of ViVo as a learning tool and as a research tool in more detail.

Pilot study spring 2008

I conducted a pilot study with 70 university students in spring 2008 to examine if students' vocabulary knowledge improved due to the use of the vocabulary trainer ViVo. Thirty-five students worked with the online vocabulary trainer ViVo, the other group of 35 students did not. Quiz test results indicated that students who had practiced regularly with the program obtained higher scores.

Furthermore, a questionnaire asked students to indicate which features they would favour in an online trainer and which learning strategies they

preferred. The majority favoured multimodal functions (i.e., sound, images, and spell-check). Mostly, they preferred repetition strategies (i.e., writing word lists, flash cards).

Following this pilot study, two questions soon emerged. First, how often should students repeat their learning material? Second, is there an optimal spacing between these practice sessions? These questions are addressed in the study presented here.

ViVo design fall 2008

This research used the web application ViVo. Its use was twofold. First, it was an online vocabulary trainer, and second, it performed as a research tracking tool.

As an online vocabulary trainer, it was based on a concept of cyclical learning and revision as discussed in the literary review (Baddeley, 1997, 2007; Ellis, 1995; Hulstijn & Laufer, 2001). It was also a learning tool that contributed to the concept of explicit vocabulary instruction as a learning process which requires *attention* (Schmidt, 2001). Its programming was informed by the research on FonFs (Laufer, 2006; Nation, 2001). As an online tool it displayed advantages that a non-electronic, handheld device such as Leitner's *Lernkartei*[25] did not provide. In their notes on Leitner's device Mondria and Mondria- De Vries (1994) recommended the design of an electronic version. They noted practicability concerns with the handheld version. At the same time, they advocated the expanded multimodal presentation options a computer version would offer.

In addition to these presentation options, ViVo offered the functionality of a research tool. Unlike the *Lernkartei*, where practice intervals were in random distribution, ViVo controlled them. Furthermore, it functioned as a tracking tool

[25] See 2.3.2.

and documented the participants' entries thus allowing for the research to be conducted in the environment of an undergraduate course setting.

For this research, ViVo was programmed to reflect two learning schedules: one with graduated intervals, the other with uniform intervals. This design is in accordance with the first research question: How will students' performances compare if one group has a schedule of graduated learning intervals and the other group has uniform ones?

ViVo as learning software

ViVo is a web-based platform specifically designed by the researcher to conduct this study. It was introduced to the students at a CALL orientation. Staff members of the CALL facility were able to provide user support. And therefore, students were advised to use ViVo on campus computers even though ViVo was accessible from any computer with an internet connection. Students could also contact the research supervisor, if they encountered problems. Furthermore, ViVo's program administrator could be contacted by email to resolve technical issues.

ViVo presented the learning material in a structured frequency. Its sequencing algorithm implemented the rehearsal schedules of graduated intervals for one group (GG) and uniform intervals for the other group (UG).

The graduated intervals were based on the *forgetting curve* (Atkinson & Shiffrin, 1968; Ebbinghaus, 1885). According to this forgetting curve, learners need to practice and review learning material within a certain time frame in order to eventually retain this knowledge in their long-term memory. This condition was met in the graduated interval schedule. Their five sessions were scheduled over the 14 days of a chapter practice in the following way:

Table 5: Graduated interval schedule

	Task	date	interval length
	distribution in class	day 1	
1.	encounter – practice and immediate review	day 1	0 interval
2.	encounter - review	day 1	½ day
3.	encounter - review	day 2	1 day interval
4.	encounter - review	day 4	2- day interval
5.	encounter - review	day 7	3- day interval
	online quiz	day 11	4-day interval
	print quiz	day 15	4-day interval

Their online quiz was scheduled on day 11 with a 4-day interval to their last review session. Their print quiz was scheduled on day 15 with an 8-day interval to their last review session and a 4-day interval to the online quiz.

The uniform schedule was informed by research on spaced learning versus massed learning. Spaced learning yields significantly higher retention rates, a phenomenon referred to as *spacing effect* (Baddeley, 1997, 2007; Hulstijn & Laufer, 2001). This group's rehearsal sessions were programmed with uniform intervals as listed below in Table 6.

Table 6: Uniform interval schedule

	Task	date	interval length
	distribution in class	day 1	
1.	encounter – practice and immediate review	day 1	0 interval
2.	encounter - review	day 3	2- day interval
3.	encounter - review	day 5	2- day interval
4.	encounter - review	day 7	2- day interval
5.	encounter - review	day 9	2- day interval
	online quiz	day 11	2- day interval
	print quiz	day 15	4- day interval

These students' online quiz was scheduled on day 11 with a 2-day interval to their last review session. Their print quiz was scheduled on day 15 with a 6-day interval to their last review session and a 4-day interval to the online quiz.

ViVo's design

ViVo's concept can be compared to a flash card system with target items grouped as pairs: L1 and L2 representation. Its functions resemble Leitner's (1972) hand computer. However, as a web application it allows for additional presentation features such as interactive tasks. The implementation of these additional features was informed by research on multimedia learning (i.e., Jones & Plass, 2002; Kim & Gilman, 2008) and that on different learning style preferences (Cohen, 2003; Oxford, 2003). Oxford (2003) described these as individual preferences along "social style dimensions", "sensory style dimensions", and "cognitive style dimensions" (p.273). Even though the added features seemed limited when compared to the *'what would be nice to have'*, they allowed to cater to a broader spectrum of learner needs. This eliminated

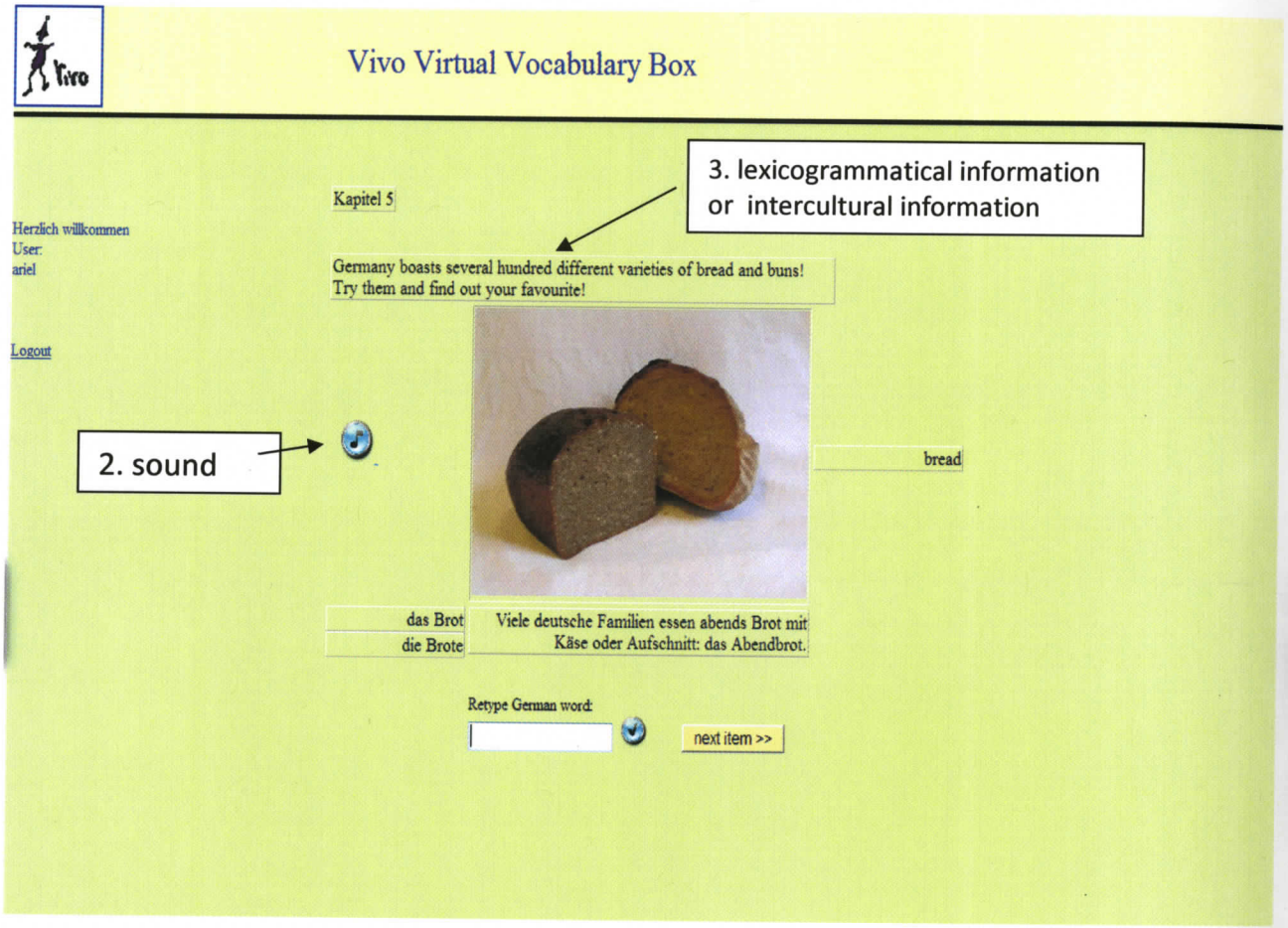
the bias a single mode might have had. Rimrott (2009) explored different presentation modes in her research on annotations that best support vocabulary learning. Her findings suggested that though a combination of options was overall better because learners fared better on annotation styles that suited their individual learning preferences. Furthermore, she found that pictorial annotations worked best for most.

Finally, the contextualization and expanded intercultural and lexicogrammatical information in ViVo corresponded to a concept of depth of vocabulary knowledge (Nation, 1990): the assumption that the acquisition of a target item is a process and encompasses the sum of *subknowledges* (i.e., register collocation, pronunciation, grammatical features, and morphological information). This concept informed ViVo's design, but because ViVo was a program for beginners' use, it represented target items in their most frequently used meaning and basic information. In addition to an L1 and L2 target item presentation, I designed ViVo with the following features:

1. image files
2. sound files
3. lexicogrammatical information
4. target language sample sentences
5. intercultural information
6. practice writing field with spell check

The screenshot below shows the "Practice" interface.

Figure 12: Graphical User Interface, ViVo in Practice Mode



Images

The version of ViVo used for this research displayed 200 images as pictures, illustrations, cartoons or clipart designs. This display was informed by the research on pictorial annotations effects (Chun & Plass, 1996; Jones, 2004; Kim & Gilman, 2008; Rimrott, 2009; Yoshii, 2006). Images are considered an effective way to improve learning of vocabulary because graphics can illustrate what the vocabulary means and often bypass an L1/ L2 translation. In combination with the sample sentence, the images used in ViVo could expand, illustrate, or complement the meaning of a target item. The word *"verlieren"* for

example is used in more than one context. It means to lose a game, but it also refers to losing items out of a pocket or basket. These semantic variants are explained in the top textbox. Then, the sample sentence depicts the first meaning (*"Die Fußballspieler verlieren das Spiel"* – the soccer players lose the game), whereas the image displays a situation of a person losing an item out of a basket.

Figure 13: ViVo image A "*verlieren*"- to lose

Often, the images used an element of humour as shown below for the target item 'which = *welcher*'.

Figure 14: ViVo image B "*welcher*"- which

Sound

The sound recordings were produced with standard High German pronunciation[26] without background noise. Native speakers' voices varied: adult female, young boy, and adult male speaker. The sound files played tribute to Baddeley's concept of the phonological loop in the working memory (Baddeley, 1997, 2007). According to his research, repeated immediate audio input aids retention. A mouse-over button provided repeated listening options and allowed students to practice pronunciation following the prompt.

Lexicogrammatical information

Examples of lexicogrammatical information were plurals and verbs with stem vowel changes. This information was explained in the top textbox and items were used accordingly in the sample sentences.

Sample sentences

Target language sample sentences were chosen according to the vocabulary proficiency level of every chapter. But, they also made use of additional cognates where applicable.

Intercultural information

The intercultural textbox drew the learners' attention to specific linguistic and cultural use of the target item. Even though the textbox entries were minimalistic and even appeared to be anecdotic, they nevertheless aimed to heighten the awareness for cultural characteristics in the L1 to L2 translation. Risager (2006) referred to this integrative interface between language and

[26] The speakers were L1 speakers of German from a town close to Hannover. The German spoken in this region is considered Standard High German.

culture as practices that should not only be studied as different conceptual views but should also be identified as crucial in the development of cultural identities. The following sample lexical entry on "bed – *das Bett*" illustrates this point:

> "Why do North Americans refer to bed size as King and Queen size whereas the same sized bed is called "*Ehebett*"- spouses' bed (literal translation) in German?"

Speculating about possible cultural concepts governing this semantic differentiation allowed learners to ponder the complexity of meaning: what it is speakers of a language actually call *bed* and to explore cultural concepts that go beyond the surface meaning.

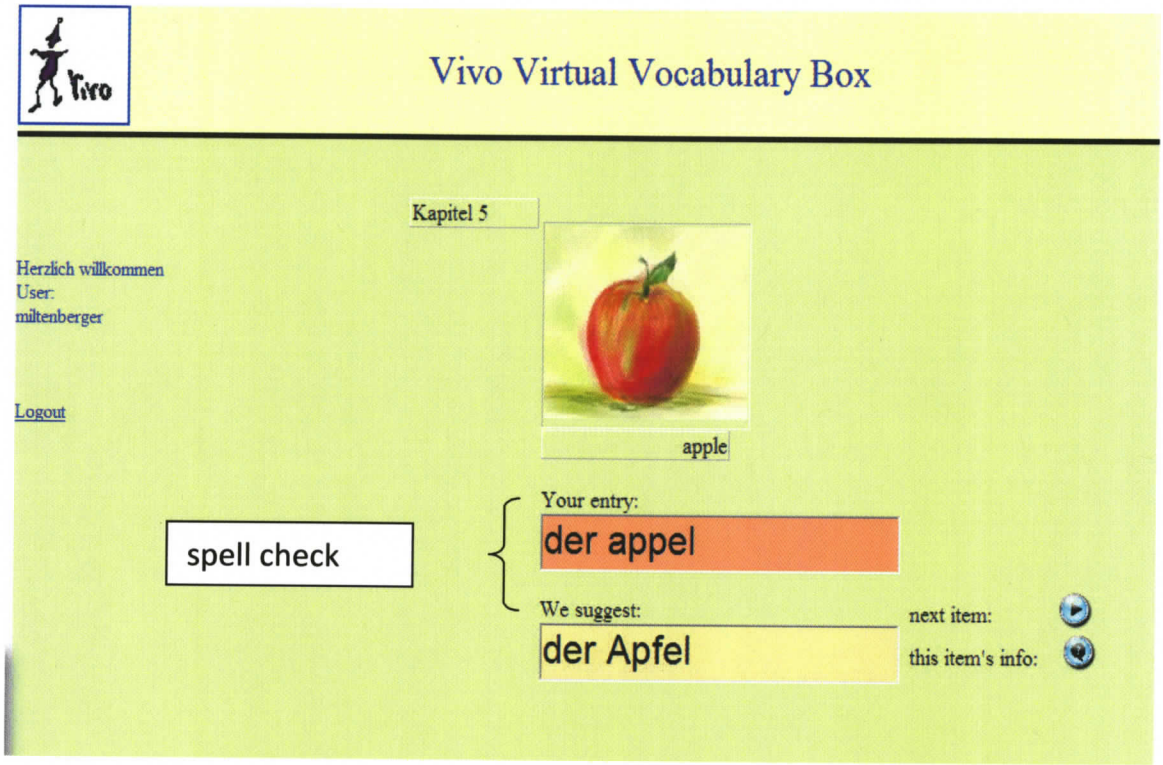
Writing practice with spell-check

Students also practiced writing the target item. Correct spelling was indicated by the background colour; wrong spelling was corrected in a prompt. This self-testing procedure was graded from mere copying of the items to production induced by an L1 prompt. Therefore, the program had two conditions: *Practice* and *Review*. Again, the programming of these two conditions was informed by research on the phenomenon of *the retrieval practice effect* (Baddeley, 1997; Ellis, 1995; Nation, 2001) or *testing effect* (Roediger & Karpicke, 2006). Self-testing with an immediate corrective feedback was used as a means to strengthen retrieval routes and thereby long-term retention.

In sum, the users could listen to sound, practice spelling, enjoy and explore picture content, obtain lexicogrammatical and intercultural information, and view sample sentences.

The *Review* interface provided the option to produce the L2 target item.

Figure 15: Graphical User Interface, ViVo in Review Mode



Again, a spell check gave immediate corrective feedback. If necessary, users could return to the *Practice* interface of the one item they were studying. However, users could not go back to the beginning of the training session. Once their session of 40 items had been completed, they received the message "You are done for today." This ensured that while students had unlimited time to study and complete their session, they would only do so according to their schedule of practice intervals.

In sum, this web application provided controlled, enriched, and multimodal training tasks for beginners.

ViVo as a research tool

All user entries were tracked and documented in ViVo's database. This included a protocol of user answers, their corrections, and user procedure (i.e., how often did they go back to the *Practice* interface when their answers were incorrect). The analyses of the user log were used to establish how often and when the GG and UG students had accessed and completed the rehearsal tasks.

First, this information was used as filter criteria. Eight student records of those who had worked with the program less than twice or not at all were removed from the data pool. Reasons for not using the program varied. Some students failed the course or dropped out late; four expressed their dislike of online resources.

Second, I used the information when students had accessed the system to create a last encounter distribution map[27] for the group of students who had only practiced in two or three sessions.

Third, ViVo user logs constituted one variable in the ANOVA data analyses as discussed in 4.3. Students had to complete five practice or review sessions per chapter.

Finally, tracking student results revealed that the 4-encounter results were very similar to those for five encounters. Therefore, students who had completed four or five encounters were grouped together, henceforth called 4/5 encounter group. A similar result could be observed for the two and three encounters. So, students who had completed two or three encounters constituted another group, henceforth referred to as 2/3 encounter group.

3.4.5 Vocabulary quizzes

Date, duration, content, and format of all vocabulary quizzes were the same for GG and UG. With the exception of the vocabulary quiz that was part of

[27] See figure in 4.3.1.

the final they were not cumulative but referred to the chapter that had just been covered in class. Online quizzes and print quizzes that were considered for the data collection of this research are described in detail below.

Online quiz

Online quizzes consisted of vocabulary items learned in the previous chapter. They were posted on the course management site Moodle and scheduled for the Friday before the print quiz was done in class. I created them as online tests using the application software HotPotatoes[28]. Their format was an L1 to L2 production of those 40 vocabulary items that had been studied with ViVo. Even though these online quizzes were only graded for completion as part of the GER100A online activity evaluation, I could download Excel tables of the grades and all student responses the course management system Moodle had compiled and use them to document students' scores.

Not all students' quizzes were used for data analyses. Tests of some students were excluded from the data analyses in accordance with the filter criteria as described above. Next, some students chose to skip this assignment; others did not complete it on the scheduled day. Finally, data from students who had not practiced for the chapter quiz using ViVo at least two times was excluded from the data analyses. Therefore, only a total of 233 online quizzes constituted the data compiled for this research part.

Print quiz

I designed the print quizzes, and the instructors approved them on their weekly meeting. The print quizzes were graded, not cumulative; and their

[28] http://hotpot.uvic.ca/

content covered the preceding chapter. Therefore, they were administered in class-time at the end of every chapter. They consisted of the parts A and B.

Part A had 20 items out of the 40 that had been studied with ViVo. These were presented in a random order to prevent a *list effect* (Nation, 2001). The task itself was an L1 to L2 translation. For this part, I established the following correction guidelines[29].

Table 7: Correction guidelines for Part A of the print quiz

0 points	.5 points	1 point
nothing was written	more than 75% of letters were correct	semantically correct choice and 100% of letters correct, AND capitalization and special characters were correct
the item written was semantically wrong	capitalization of nouns was omitted special characters were omitted or wrong	

Part B consisted of 10 items. I chose these from the active–passive vocabulary items of the chapter that had not been studied with ViVo. The test tasks for these items were designed as either receptive multiple choice questions or open-ended questions. The scoring protocol for multiple choice type questions (i.e., Chapter 4, print quiz, No. 26: "to come by" means: a. *zurückkommen* b. *mitkommen* c. *vorbeikommen* d. *beikommen*) was 0 points for wrong answers, or one point for the correct answer. The open-ended questions (i.e., Chapter 5, print quiz, No. 28/29/30: Name three pieces of clothing in German you typically wear in winter) were graded as described in Table 8.

[29] This simplified scoring protocol was informed by Keating (2008).

The students received an overall mark for Part A and Part B. For research purposes, however, the grades of these parts were also calculated separately. The results of Part A and Part B could then be compared in an ANOVA data analysis to establish whether using a vocabulary trainer would indeed improve retention and higher scores could be attributed to its use. Next, the results of Part A were used as the dependant variable in the ANOVA data analyses of both research questions.

Again, not all student quizzes were used in the data analyses. Some were excluded in accordance with the filter criteria described in this section. Also, the quizzes of those students who had missed their written test on the test date were excluded. Furthermore, the quizzes of those students who had not studied at least two times with the vocabulary trainer were excluded from the data as well. Therefore, a total of 361 quizzes constituted the data compiled for this research.

3.5 Statistical analyses

A series of One-Way analyses of variance (ANOVA) was used to set basic parameters on behalf of the efficiency of the research tool and to answer the research questions. This procedure is described below.

3.5.1 Efficiency of online tool ViVo

In addition to the pilot study described above, I chose ANOVA analyses to test the efficiency of the online vocabulary trainer as a basic parameter of this study. The conditions *practice with ViVo* and *non-practice with ViVo* were compared. As described earlier, the print quiz consisted of two parts. Part A displayed words practiced with ViVo, whereas Part B contained words that had not been studied with ViVo. Therefore, Part A test scores were compared with

Part B test scores. Part A and B constituted the independent variable; the test scores were the dependent variable. My hypothesis was that items studied with the online trainer would demonstrate a higher retention rate than those that had been studied traditionally and without the aid of a training program. The table below provides an overview of the data that was used in the statistical analyses comparing print quiz Part A and Part B.

Table 8: Comparison of print quiz results Part A and B

Participants	Interval practice groups	Chapters
all	UG and GG	all chapters
4/5 encounters	UG and GG	all chapters
	GG	all chapters
	UG	all chapters
2/3 encounters	UG and GG	all chapters
	GG	all chapters
	UG	all chapters

3.5.2 Interval variations: first research question

For the first research question, the test conditions of interval variations constituted the independent variables, the student test scores were the dependent variables. First, I compared data from all tests of all participants who had completed 4/5 encounters. The data from students who had completed 2/3 encounters were excluded from the analyses of this research question because their practice schedule did not comply with the test conditions of graduated or uniform intervals. Their intervals can be best described as random intervals because students chose to skip encounters. The timing of these encounters was therefore a random distribution. For example, students could have practiced on

their first, second, and third encounter (1-2-3), or their third, fourth, and fifth encounter (3-4-5).[30] Therefore, I chose not to include this random distribution data for this research because test group sizes would have been too small. More data must be compiled to reach comparable test group sizes for further quantitative research[31].

Next, I examined the data more closely with regard to tests (print quiz and online quiz) and finally, I examined chapter differences. The table below provides an overview of the tests and their grades that were used in the statistical procedures.

Table 9: Interval spacing, overview of the tests used in the statistical analyses

Participants	Tests
all 4/5 encounters	all
	all online quizzes
	all print quizzes
	individual chapters 1 – 5 online quizzes
	individual chapters 1 – 5 print quizzes

3.5.3 Frequency of encounters: second research question

For the second research question, I evaluated the test score results comparing the 4/5 encounter group with the 2/3 encounter group. This time, the two encounter groups constituted the independent variable and the test scores were the dependent variable. Again, I examined this data more closely with regard to test conditions (uniform interval and graduated interval) and tests (print quiz or online quiz). I did not examine individual chapters because the number of participants was too low in the 2/3 encounter test condition. Some

[30] Other encounter variations: 1-2-4; 1-2-5; 2-3-4; 2-3-5.

[31] Present studies in spring 2009 and in fall 2009 also compile this data. This study's data of the 2/3 encounters can then be included in further research.

chapters had five or less participants (n≤ 5). The table below provides an overview of the test data that was used in the statistical analyses for the second research question.

Table 10: Interval frequency, overview of the tests used in the statistical analyses

Participants	Tests
graduated and uniform interval groups	all
	all online quizzes
	all print quizzes
uniform interval group	all online quizzes
	all print quizzes
graduated interval group	all online quizzes
	all print quizzes

The 2/3 encounter group practiced at three respectively two scheduled practices. The time lapse between their individual last encounter and their tests was defined by the encounter they chose to do last. I therefore mapped the distribution of the last practice for 2/3 encounters. In a more qualitative research approach, I analysed the students' responses to Questionnaire B questions 11, 12, and 13. These addressed a self-evaluation of students' needs for more or less encounters.

4. Data analyses and results

The purpose of this chapter is to present the data analyses results. It is organized in the following way. First, all validity issues are addressed: homogeneity of the test groups, effectiveness of the online tool ViVo, and possible bias by incidental encounters. Next, the interval variation data analyses and its results (first research question) are introduced. Then, findings on behalf of score differences between students who had studied 2/3 times and those who had studied 4/5 times (second research question) are presented.

4.1 Validity

Not all test conditions can be controlled in a field experiment. The following presents how the analyses of data from Questionnaire A, Questionnaire B, textbook frequency scan, and the ViVo user log provided a validity framework. First, these helped to assess the possibility of biased results due to uncontrolled exposure of the target items; second, to establish a homogeneous group of participants; and third, to establish the effectiveness of the learning tool.

4.1.1 Incidental encounters

This part explored whether or not incidental encounters influenced the learning outcome. A score analysis indicated that there were no significant differences between results for items that had been encountered often in the chapter and those that had been encountered less often. The two tables below present the average scores for the randomly chosen 6 vocabulary items. The first table indicates print quiz scores; the second refers to the online quiz.

Table 11: Incidental encounters, print quiz results

word	Buch	kaufen	Zucker	wie	Wohnung	mit
frequency[a]	(7)	(8)	(2)	(53)	(41)	(28)
L1	*book*	*buy*	*sugar*	*how*	*apartment*	*with*
UG	89.01	98.07	93.91	89.01	93.73	90.22
GG	94.38	98.38	97.64	91.24	94.87	83.97
total	91.69	98.22	95.77	90.12	94.30	87.09

Note. a. Frequency in the textbook *Deutsch NaKlar.*

The print test scores for the frequent items (90.12 %; 94.30 %; 87.09 %) were not higher than those of the less frequent ones (91.69 %; 98.22 %; 95.77 %). On the contrary, they were even lower. The online quiz scores revealed similar results. Again, less frequent words showed even slightly better scores.

Table 12: Incidental encounters, online quiz results

word	Buch	kaufen	Zucker	wie	Wohnung	mit
frequency	(7)	(8)	(2)	(53)	(41)	(28)
L1	*book*	*buy*	*sugar*	*how*	*apartment*	*with*
UG	90	89	95	97	84	93
GG	87	88	91	85	82	89
total	88.50	88.50	93	91	83	91

A closer look at the errors produced for these items did not indicate that frequent items differed. Capitalization, gender, spelling, and word meaning errors appeared for both frequent and less frequent items.[32]

In conclusion, students did not do better on either print quizzes or online quizzes.

[32] See Appendix F for the online quiz error protocol of these items.

4.1.2 Homogeneous groups

Data sets of 86 students from the 117 registered in GER100A sections were incorporated in this study. Seventeen students did not match the profile of a German language beginner because they had either had prior exposure to German language or significant German heritage contacts. The data of eight students could not be used because they had not worked with the online vocabulary trainer ViVo, or they had missed vocabulary quizzes. Another group of six students was excluded because they had expressed their preference for an extensive use of other flashcard systems. These filter criteria were therefore able to establish two test groups that compared in their social and educational profiles.

4.1.3 Effectiveness of vocabulary trainer

Another precondition of the research was concerned with the effectiveness of the vocabulary trainer itself. Two aspects: learning gain and user friendliness were examined.

The first aspect considered a measureable learning gain that could be attributed to its use. ANOVA analyses, comparing test scores for vocabulary items studied with ViVo with those not studied with ViVo, provided evidence for the students' improvement. These are described below in more detail. For the second aspect, user friendliness, student answers and comments as documented on their Questionnaire B were explored and used for a qualitative evaluation (see section 5.1).

Comparison of print quiz results

A comparison of the print quiz results Part A and Part B for all chapters answered the question about ViVo's effectiveness favourably. Both interval test groups benefited significantly from the use of the online trainer. Furthermore,

students studying in 2/3 practice sessions as well as those studying 4/5 times did much better on their quiz part where they had studied the target items using ViVo.

As mentioned before, the print quizzes were made up of two parts. Part A constituted the ViVo practiced test corpora, Part B consisted of vocabulary that students had not learned with ViVo. Considering that Part A's task was a productive assignment whereas Part B's was a receptive one, this outcome is even more astonishing. Usually, students do better on tests that assess receptive skills (Ellis & Beaton, 1993; Nation, 2001; Schmitt, 2000, 2002). But in this study, students who had used the online trainer were doing much better on their quiz Part A than on Part B.

These findings are presented in detail below in the following order. The print quiz results of the two test parts were compared in One-Way ANOVA analyses. First, the overall results of participants who had practiced in 2/3 encounters and 4/5 encounters were examined. Next, possible differences between the test conditions graduated intervals and uniform intervals were examined.

Results of 2/3 encounters and 4/5 encounters

The table below shows a summary of the ANOVA statistics. It provides grade averages and significance factors.

Table 13: Effectiveness of ViVo - comparing print quiz Part A and Part B

N	Comparing encounters	*M* Part A	*M* Part B	*p*
$n^a = 118$ $n^b = 118$	UG and GG with 2/3 encounters	89.47	79.40	*.000*
$n^a = 243$ $n^b = 243$	UG and GG with 4/5 encounters	93.81	79.34	*.000*

The results show that students who had practiced 2/3 times with ViVo did significantly better ($p = .000$) on their Part A than on Part B. They achieved a grade average of 89.47 % on Part A and only 79.40 % on Part B. Similar findings applied to the group of students who had studied with ViVo during 4/5 practice sessions. They achieved a 93.81 % grade average on Part A. Their Part B scored a 79.34 % average and thus led to a significant difference in the learning outcome ($p = .000$).

Students who had learned with the ViVo vocabulary trainer fared better on these quiz parts. Additionally, the fact that the grade average for the 4/5 encounters was higher than that for the 2/3 encounters indicated that more practice would lead to better results. This will be discussed further on as one aspect of the second research question: *How many encounters are best for optimal results*. Interestingly, the results for test parts not practiced with ViVo were the same for both groups (79.34 % and 79.40 %). This, too, suggested that differences in learning gain could therefore be attributed to vocabulary practice with the online trainer.

Results for graduated intervals compared to uniform intervals

Students' performance in the two test groups with varying test conditions, namely graduated spaced intervals between practice sessions and uniform spaced sessions, did not show any difference with regard to the effectiveness of the vocabulary trainer. Both groups benefited equally well from the trainer. The table below shows these results in detail.

Table 14: Effectiveness of ViVo for uniform and graduated test condition

N	comparing intervals	*M* Part A	*M* Part B	*p*
$n^a = 120$ $n^b = 120$	4/5 encounters UG	94.93	81.12	*.000*
$n^a = 123$ $n^b = 123$	4/5 encounters GG	92.72	77.60	*.000*
$n^a = 65$ $n^b = 65$	2/3 encounters UG	89.38	79.23	*.000*
$n^a = 53$ $n^b = 53$	2/3 encounters GG	89.57	79.62	*.002*

Students practicing 4/5 sessions on uniform intervals had an average 94.93 % grade on Part A of their quiz and 81.12 % on Part B. This constitutes a significant increase ($p = .000$) in student performance. Students practicing 4/5 sessions on graduated intervals demonstrated a performance of 92.72 % on Part A and 77.60 % on Part B. Again, these results indicate that ViVo practice contributed significantly ($p = .000$) to the learning gain. However, it did not matter whether students had practiced on a graduated schedule or a uniform schedule. I observed similar results for practice encounters of 2/3 sessions. Students on a uniform schedule had significantly ($p = .000$) high results on Part A of their quiz ($M = 89.38$ %) and lower ones on their Part B ($M = 79.23$ %). This applied to students on a graduated schedule, too. They performed well on Part A ($M = 89.57$ %) and less well on Part B ($M = 79.62$ %). Again, this significant difference ($p = .002$) could be attributed to the fact that students studied Part A with the vocabulary trainer but not Part B.

User friendliness and acceptance of online tool

As described in 3.4.3, Questionnaire B provided means to elicit information on behalf of students' acceptance of the program. Question 3 addressed preference of online tools versus conventional learning methods.

Question 5 required students to check those electronic resources they felt comfortable using. Finally, open format Questions 14 and 15 asked for favourite parts about ViVo and suggestions for improvement. The following results are presented based on Questions 3 and 5. Quotes in this text are used to illustrate students' opinions and provide evidence for my claims.

Those students, who answered Question 3, favoured an online tool for vocabulary learning (82.55 %). Only 13.95 % preferred other methods and 3.48 % did not answer this question.

Table 15: Student responses- Questionnaire B 3

N	preferred online tool in %	preferred other methods in %	question not answered in %
GG and UG n= 86	82.55	13.95	3.48

Looking at these numbers in more detail, 86.84 % of the GG and 84.44 % of the UG were in favour of the online tool. Three students did not answer this question.

Table 16: UG and GG responses - Questionnaire B 3

N = 83	preferred online tool in %	preferred other methods in %
GG n= 38	86.84	13.15
UG n= 45	84.44	15

Question 5 requested students to check mark those IT devices they felt comfortable using. Ninety-one percent of the students chose ViVo as easy to use.

In general, the acceptance of the program was positive. The following student comments[33], taken from the answers 84 students provided, exemplify this point. They were typical answers to Question 14: *What was your favourite part about ViVo*?

Table 17: Student comments - Questionnaire B 14

student	1	"Repetition and due dates"
student	6	"convenience, the repetition really helped me to remember."
student	8	"the audio"
student	79	"Being able to have all the aspects of learning (listening, reading, typing the words) in one place"
student	15	"the regular feedback and ability to immediately review"
student	20	"the images with the pictures, which made it easier to remember German meanings"
student	81	"It works! If you could remember to do it, you could get 100% on the test."
student	26	"ViVo was a constant refresher that helped me to study more regularly instead of cramming before a test."
student	70	"Besides the funny pictures? The constancy of it."
student	54	"It provided a systematic, organized method of studying, so I wasn't overwhelmed by the sheer weight of words."
student	32	"getting good marks on quizzes"

[33] Spelling or grammar was not corrected.

Interestingly, students' comments covered a wide range of features they liked most. They addressed i.e., the multimodal design, the easy access, and the structured learning process.

Question 15 asked students for their input on how to improve a vocabulary online trainer. Ninety-three students answered this question (UG = 49; GG = 44). I grouped these answers as follows:

a. addressing frequency (more or less practice)
b. requesting more flexible schedules
c. suggesting innovative features
d. commenting on user issues and technical glitches
e. nothing to report

In particular, answers to a, b, and c were of interest to this study. Fourteen comments addressed frequency (UG = 10; GG = 4).The following samples of their comments reflect this.

Table 18: Student comments - Questionnaire B 15a

student	8	"give more opportunities to practice."
student	27	" – make it available every day"
student	58	"There's not too much that I would do to improve Vivo except to maybe allow more access to the program."

For 25 students, though, not frequency was the main concern but flexibility. They commented on the flexibility to access the program. The following samples of their comments illustrate this point.

Table 19: Student comments - Questionnaire B 15b

student	55	"-> more flexibility with the schedule i.e., times and days"
student	74	"easier to remember schedule"
student	38	"- Have no set days, just make it mandatory to complete 5 sessions in a time period."
student	11	"Allow the schedule to be flexible, or have it open 24/7, with a mandatory period closed after a session."
student	22	"Making it more available, rather than just on certain days, for those who don't have as much free time."

Another group of 28 students put interesting ideas forward to include in the program. Their suggestions ranged from multimodal feature[34] add-ons (i.e., games, video, contextualized input on sentence level) to methodological issues (i.e., lexical presentation in random word order, cumulative testing, categorization), and technical improvements (i.e., special character key pads, optimization of spell checker). These multifaceted contributions to a future design mirrored the interest students showed in the product and its implementation.

4.1.4 Summary

In conclusion, I established the following validity framework. First, I argued that the validity of this study was not influenced by the frequency of the target items in the textbook. Second, I presented how two homogeneous test groups were set up. Third, I explained how the online vocabulary trainer contributed significantly to the learning gain and could therefore be implemented as a vocabulary learning tool in this research.

[34] I did not count features that were requested yet were already implemented in the program (i.e., sound).

The following paragraphs will now report on the findings for the two research questions.

4.2 Interval schedules

This paragraph presents the findings in the following order. First, the overall interval schedule results are discussed. They are followed by the detailed description of online quiz results and print quiz results. Next, I report on the findings of every chapter, again with regard to online quiz results followed by print quiz results.

The results of the study provided intriguing answers to the research question:

How do students retain vocabulary best: following a practice schedule of uniform intervals or following a schedule of graduated intervals?

According to the research findings of this study, it did not matter which practice schedule students followed. A graduated spaced practice schedule elicited the same learning gain as a uniform spaced practice schedule.

One-Way analyses of variance (ANOVA), calculated on the basis of all participants (UG n = 202; GG n = 212) and all 10 quizzes, revealed no significant differences ($p = .183$) between the two test conditions.

Table 20: Overall interval schedule results —all tests, all participants

N= 414	*M* GG	*SD* GG	*M* UG	*SD* UG	*p*
UG n= 202 GG n= 212	92.09	9.71	93.43	10.78	*.183*

Students studying with a graduated schedule obtained a mean average of 92.09 % on their tests. Students following uniform spaced study intervals

achieved a slightly higher 93.43 % grade average on their tests. The standard deviation showed no big differences: *SD* = 9.71 (GG) and *SD* = 10.78 (UG). Though the mean average of the uniform condition was higher, this increase was not significant.

The following analyses now break these test findings down into a more detailed account. The first part provides the online quiz results for both conditions. The second part presents the results for the print quizzes, and the third part examines possible differences within print quiz chapters.

4.2.1. Online quiz results

These analyses looked at score differences of all five online quizzes that could be attributed to uniform versus graduated interval conditions. Again, a series of One-Way ANOVA analyses revealed no significant differences between the two test conditions.

Table 21: Interval schedules, online quiz results

N[a]	encounters	*M* GG	*SD* GG	*M* UG	*SD* UG	*p*
UG n= 82 GG n= 89	4/5	91.22	8.25	91.24	14.08	*.991*

Note. [a] Students' online quizzes that were not completed on the due date were not included (see 3.4.5).

The group that had studied 4/5 times showed a significance factor of *p = .991*. The grade averages were high and indicated that using the program led to better results. The 4/5 encounter group on a graduated schedule scored a 91.22 % average; the students on a uniform spaced schedule achieved the slightly higher result of 91.24 %. In sum, graduated and uniform interval conditions did not show significant differences in their online test scores.

4.2.2. Print quiz results

A more specific aspect of the research question was: Could differences between graduated condition and uniform condition be identified in the print quiz? A series of One-Way ANOVA analyses revealed no significant differences between the two test conditions. Students in the UG performed as well as students in the GG.

Table 22: Interval schedules, print quiz results

N[a]	encounters	*M* GG	*SD* GG	*M* UG	*SD* UG	*p*
UG n= 120 GG n= 123	4/5	92.72	10.63	94.93	7.48	*.062*

Note. [a] Students' print quizzes not written on the test date were not included (see 3.4.5).

Differences in print test scores were not significant with regard to the test condition uniform and graduated intervals. However, the *p* value (*p = .062*) was close to the significance factor. A closer look at the standard deviation revealed that the GG results were higher (*SD* GG = 10.63 to *SD* UG = 7.48). This factor contributed to the close significance level. Nevertheless, the average test scores showed that, even though not significant, UG students performed better on their print quiz.

In sum, graduated intervals and uniform interval conditions did not yield significant differences in the print test scores. Students who had practiced 4/5 times and learned on a graduated schedule did not perform better than those students who had learned 4/5 times according to a uniform practice schedule.

4.2.3 Chapter differences

A further series of One-Way ANOVA analyses examined the test scores obtained in the chapters. Again, the test condition graduated interval was compared to the test condition uniform interval. Accordingly, I set up the following series for print quizzes (4/5 encounters) and online quizzes (4/5 encounters) comparing the interval conditions:

graduated interval		uniform interval
chapter 1	to	chapter 1
chapter 2	to	chapter 2
chapter 3	to	chapter 3
chapter 4	to	chapter 4
chapter 5	to	chapter 5

Chapter differences in online quiz results

This section looked at the scores of the online quizzes comparing the chapter results with each other. Only data of those students who passed the filter criteria screening were included (see 3.4 for details). Graduated interval condition and uniform interval condition were compared. None of the chapters revealed a significant difference between graduated condition and uniform condition. However, three chapters measured a slightly higher mean for the uniform condition. The chapters are described in detail below.

Table 23: Interval schedules, online quiz chapter results

N	chapter	*M* GG	*SD* GG	*M* UG	*SD* UG	*p*
UG = 21 GG = 25	1	83.60	19.69	88.23	20.47	*.439*
UG = 12 GG = 13	2	93.15	5.42	95.91	1.92	*.109*
UG = 17 GG = 16	3	94.50	8.47	96.05	5.91	*.542*
UG = 17 GG = 18	4	92.00	7.37	88.05	20.84	*.456*
UG = 15 GG = 17	5	89.35	5.55	89.26	5.18	*.964*

Chapter 1 analysis showed a mean of 83.60 % for the graduated condition and 88.23 % for uniform. This led to $p = .439$. The standard deviation values are high because minimum scores were low (UG = 20 % and GG = 11 %). Chapter 2 did not demonstrate large differences between average graduated condition results (93.15 %) and uniform average scores (95.91 %), but differences were not significant ($p = .109$) even though the uniform average scores were slightly higher. Chapter 3 analysis revealed no significance with $p = .542$ and a graduated condition mean of 94.50 % versus a uniform condition mean of 96.05 %. The chapter 4 analysis showed a mean of 92.00 % for the GG and 88.05 % for the UG ($p = .456$). High differences in the standard deviation may have accounted for the fact that, unlike in other chapters, the UG average was lower here than the GG one. The UG had a minimum score of 17 % and a high standard deviation ($SD = 20.84$), and the GG had a minimum score of 73 % and a low standard

deviation value ($SD = 7.37$)[35]. Finally, chapter 5 revealed no significant difference ($p = .964$) between the two test conditions of graduated intervals versus uniform ones. Mean averages (UG = 89.26 %; GG = 89.35 %) and minimum values (UG = 77 %; GG = 74 %) were relatively high, and standard deviation values were low and close. This corresponded to the hypothesis that students made conscious choices which chapter quizzes they practiced for completion and which they used as challenges to test themselves. The chapter 5 quiz was very close to the final exam. Very likely, they tried to do as best as possible in order to test their proficiency level.

In summary, none of the chapter results yielded a significant difference between the two test conditions: graduated and uniform intervals. However, as observed before, the UG outperformed the GG and scored higher mean averages on most chapter quizzes. Standard deviation values were generally much higher compared to the print quizzes.

Chapter differences in print quiz results

For this series of ANOVA analyses, I looked at student scores of every chapter comparing graduated interval condition with uniform interval condition for those students who had completed 4/5 practice sessions per chapter. None of the chapters revealed a significant difference between graduated condition and uniform condition. However, all chapters measured a slightly higher mean for the uniform condition. The chapters are described in detail below.

---

[35] See discussion on extreme values and high standard deviation of the online quizzes in 5.2.2 *The human factor.*

Table 24: Interval schedules, print quiz chapter results

N	chapter	*M* GG	*SD* GG	*M* UG	*SD* UG	*p*
UG = 28 GG = 29	1	94.74	10.55	97.23	4.63	*.256*
UG = 29 GG = 30	2	93.25	8.95	95.25	7.59	*.358*
UG = 20 GG = 23	3	92.82	11.73	94.00	9.01	*.718*
UG = 22 GG = 21	4	91.19	13.31	95.56	6.07	*.170*
UG = 21 GG = 20	5	90.50	8.90	91.66	9.36	*.685*

Chapter 1 analysis showed a mean of 94.74 % for the graduated condition and 97.23 % for uniform. This led to $p = .256$. Chapter 2 demonstrated similar numbers with graduated condition results at 93.25 % and uniform at 95.25 % ($p = .358$). Chapter 3 analysis revealed no significance with $p = .718$ and a graduated condition mean of 92.82 % versus a uniform condition mean of 94.00 %. The chapter 4 analysis showed a mean of 91.19 % for the GG and 95.56 % for the UG ($p = .170$). The low *p*-value could be attributed to the differences in standard deviation values. Finally, chapter 5 followed the same pattern of results with a mean of 90.50 % for the GG and the slightly higher results of 91.66 % for the UG. Again, these numbers were not significant with $p = .685$.

In sum, graduated intervals and uniform interval conditions did not yield significant differences when comparing the print test scores of individual chapters. However, all five chapters showed slightly higher means for the uniform condition.

4.2.4 Summary

This chapter provided the results of data analyses that answered to the first research question. Students' scores following a practice schedule of uniform intervals did not differ significantly from those of students who were following a schedule of graduated intervals. This was true for both print quiz results and online quiz results with 4/5 encounters. Furthermore, detailed analyses for all chapters did not show significant differences between the two test conditions in the chapter results either. Therefore, no spacing interval type took precedent over the other. Students learned well under the condition of a graduated schedule, and they fared well under a uniform practice schedule. However, most test results for the uniform test condition were slightly higher than those of the graduated test condition, but—these differences were not significant.

4.3 Practice frequency

The second research question looked at the number of practice encounters more closely. Frequency, interval schedule, desired proficiency level, and retention robustness form a quadrangle of interdependent processes. It therefore also matters how often students study.

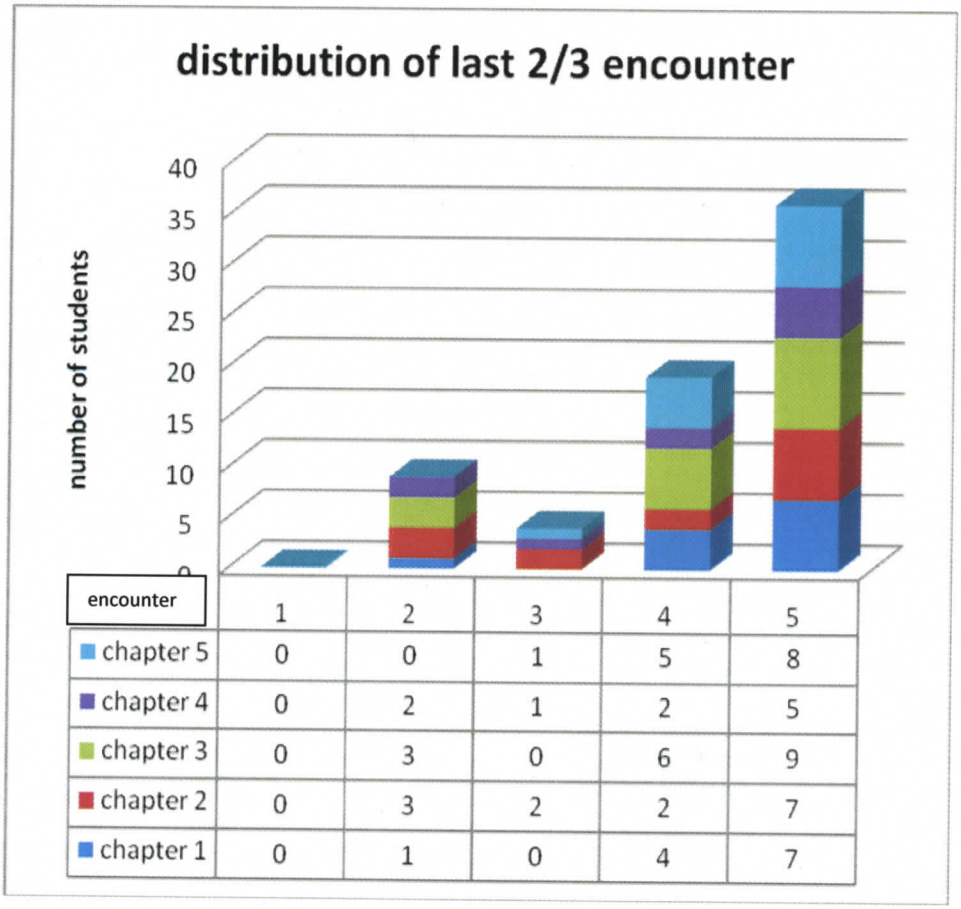
ViVo's user log compiled the information how often students had studied in its user logs and tracked which ones they had completed. In particular, this was of importance for the analyses of the 2/3 encounter distribution.

4.3.1 Last practice distribution of 2/3 encounter group

In addition to a count of the times students had accessed the vocabulary trainer, I used the ViVo user logs to obtain a distribution map of the practice sessions. It showed when students of the 2/3 encounter group chose to practice. An analysis of this distribution allowed me to identify the last time they had

practiced before their quizzes. This distribution is presented in the table below. Interestingly, the majority of students (36 out of 68) studying 2/3 times completed their last practice at the last possible session prior to the quizzes. This showed that most 2/3 encounters and all 4/5 encounters had the same interval between last encounter and the online quiz. For the UG this was day 9 with a 2-day interval to the online quiz and a 6-day interval to the print quiz. For the GG this was day 7 with a 4-day interval to the online quiz and an 8-day interval to the print quiz. This proximity to the quizzes had to be taken into consideration because it could have influenced the robustness of the learned material.

Figure 16: Distribution of last encounter



encounter	1	2	3	4	5
chapter 5	0	0	1	5	8
chapter 4	0	2	1	2	5
chapter 3	0	3	0	6	9
chapter 2	0	3	2	2	7
chapter 1	0	1	0	4	7

4.3.2 Results of 2/3 encounters versus 4/5 encounters

Test score analyses using One-Way ANOVA compared the print and online scores of students who had learned with ViVo in 4/5 practice sessions with those of students who had completed 2/3 sessions.

The One-Way ANOVA analyses calculated on the basis of all quizzes, all chapters, and both UG and GG revealed a significant difference *(p = .000)* between the two encounter groups. The 2/3 encounter group obtained a mean average of 88.59 %, the 4/5 encounter group scored 92.75 %.

Table 25: Overall interval frequency results—all quizzes, all participants, all chapters

N= 595	2/3 encounter		4/5 encounter		
	M	***SD***	***M***	***SD***	***p***
2/3 n= 181 4/5 n= 414	88.59	15.35	92.75	10.26	*.000*

However, looking at the results in more detail portrayed a more differentiated picture. The following analyses therefore break those findings down into a more detailed account looking first at online quiz results, their differences found in graduated and uniform condition and then examining print quiz results. These, too, are viewed in their graduated and uniform condition.

Online quiz results

Based on online quizzes these analyses looked at score differences between the two encounter groups. Comparing the data for graduated and uniform intervals and all chapters, the differences between 2/3 encounters and 4/5 encounters are significant *(p = .037).* The test score results of the

4/5 encounter group were higher than those of the 2/3 encounter group (4/5 *M* = 91.23 %; 2/3 *M* = 86.96 %). When comparing the variables under graduated and uniform interval condition, neither condition revealed significant differences (UG *p* = *.108*; GG *p* = *.201*), even though the mean averages for 4/5 encounters were higher (UG = 91.24 %; GG = 91.22 %) in both conditions than those of the 2/3 encounters (UG = 85.60 %; GG = 88.67 %).

Table 26: Interval frequency, online quiz results

		2/3 encounter		4/5 encounter		
N		***M***	***SD***	***M***	***SD***	***p***
2/3 n= 63 4/5 n= 171	all chapters, UG and GG	86.96	18.82	91.23	11.39	*.037*
2/3 n= 35 4/5 n= 82	all chapters, UG	85.60	23.16	91.24	14.08	*.108*
2/3 n= 28 4/5 n= 89	all chapters, GG	88.67	11.52	91.22	8.25	*.201*

Print quiz results

Analyses comparing 2/3 and 4/5 encounters for the print quizzes revealed similar results. Again, differences were significant for the total of UG and GG (*p* = *.000*).

Based on all chapter print quizzes of UG and GG, these analyses examined test result differences between the two encounter conditions 2/3 and 4/5. UG and GG students who had studied in 4/5 encounters did much better on their print quizzes with an average of 93.81 %, whereas those students who had studied in 2/3 encounters scored a mean of 89.47 %.

Table 27: Interval frequency print quiz results —all chapters, all participants

	2/3 encounter		4/5 encounter		
N[a]= 361	***M***	***SD***	***M***	***SD***	***p***
2/3 n= 118 4/5 n= 243	89.47	13.13	93.81	9.26	*.000*

Note [a] UG and GG for all chapters

Yet, a closer look at the two conditions UG and GG again, revealed more differentiated results. Whereas the UG showed significant differences between the two test conditions 2/3 and 4/5 ($p = .000$), the GG did not ($p = .116$), even though the mean averages of both UG and GG were higher for 4/5 encounters. The UG showed a mean average for 2/3 encounters of 89.38 % and 94.93 % for 4/5 encounters. The GG 4/5 group average was higher (92.72 %) than that of the 2/3 group (89.57 %), but the ANOVA analysis did not mark this difference as significant.

Table 28: Interval frequency print quiz results—UG and GG

		2/3 encounter		4/5 encounter		
N		***M***	***SD***	***M***	***SD***	***p***
2/3 n= 65 4/5 n= 120	all chapters UG	89.38	11.49	94.93	7.48	*.000*
2/3 n= 53 4/5 n= 123	all chapters GG	89.57	15.02	92.72	10.63	*.116*

In conclusion, the majority of students who had studied their vocabulary in 4/5 encounters did better on their print quizzes. Mean averages were higher for all of them. Results for the UG were statistically significant. Interestingly, these

test results corresponded to students' perceived need for 4 or 5 practice sessions as the qualitative data described below demonstrates.

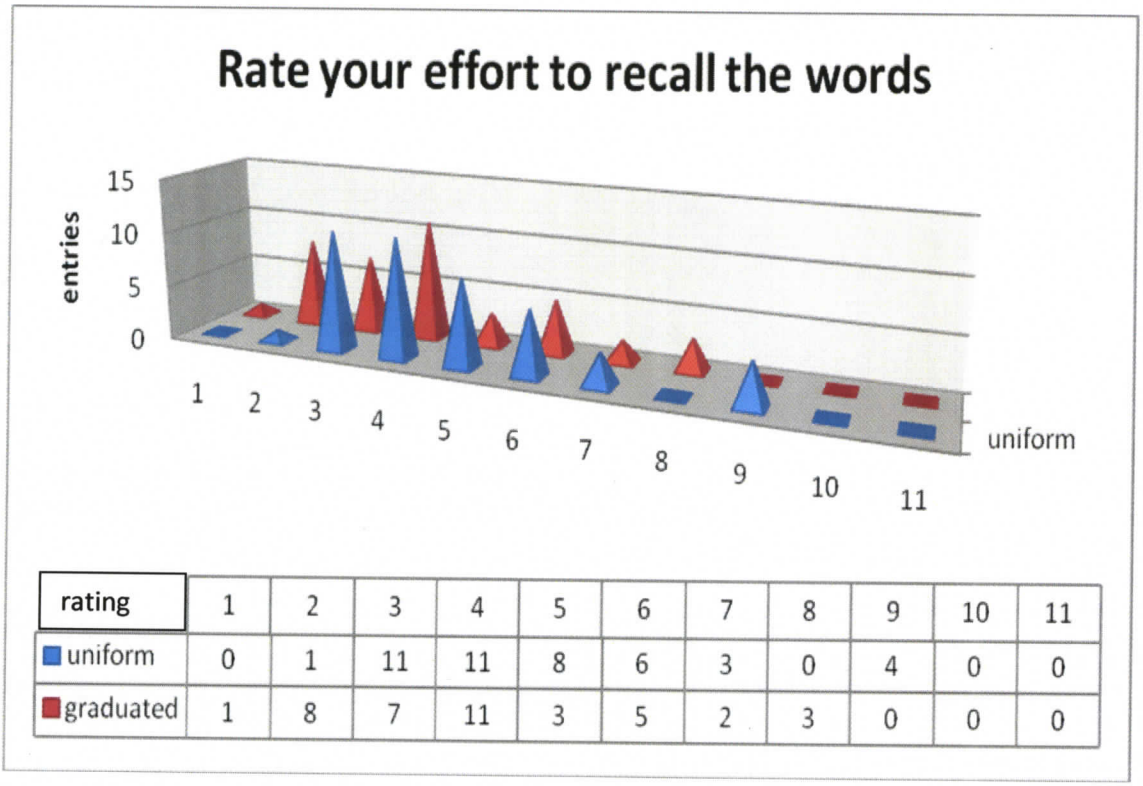
4.3.3 Student self-evaluation – Questionnaire B

In Questionnaire B, students reported on how many encounters they felt were necessary and how difficult they perceived the learning task. In particular, Question 11 addressed the ease students expressed when remembering the learnt material from one encounter to the next. Question 12 tracked reasons for having skipped encounters and Question 13 addressed possible circumstances which legitimated the limitation of encounters. Their results are described in detail below.

Question 11

Students rated their ease remembering the practiced vocabulary items from one encounter to the next on a scale from 1 (it was easy, I remembered all) to 11 (It was difficult, I had to relearn most).The figure below shows how they estimated their effort to recall the learning material. First, it shows that most students felt they had remembered enough to make the learning task easy to accomplish. The upper third on the scale reports only few entries (7 in total), whereas the lower third holds significantly more entries (50 in total). Second, apparently, the uniform group felt they needed to put more learning effort to the task. Their lower third claims only 23 entries compared to the 27 entries in the GG's lower third.

Figure 17: Questionnaire B-11, student self-evaluation

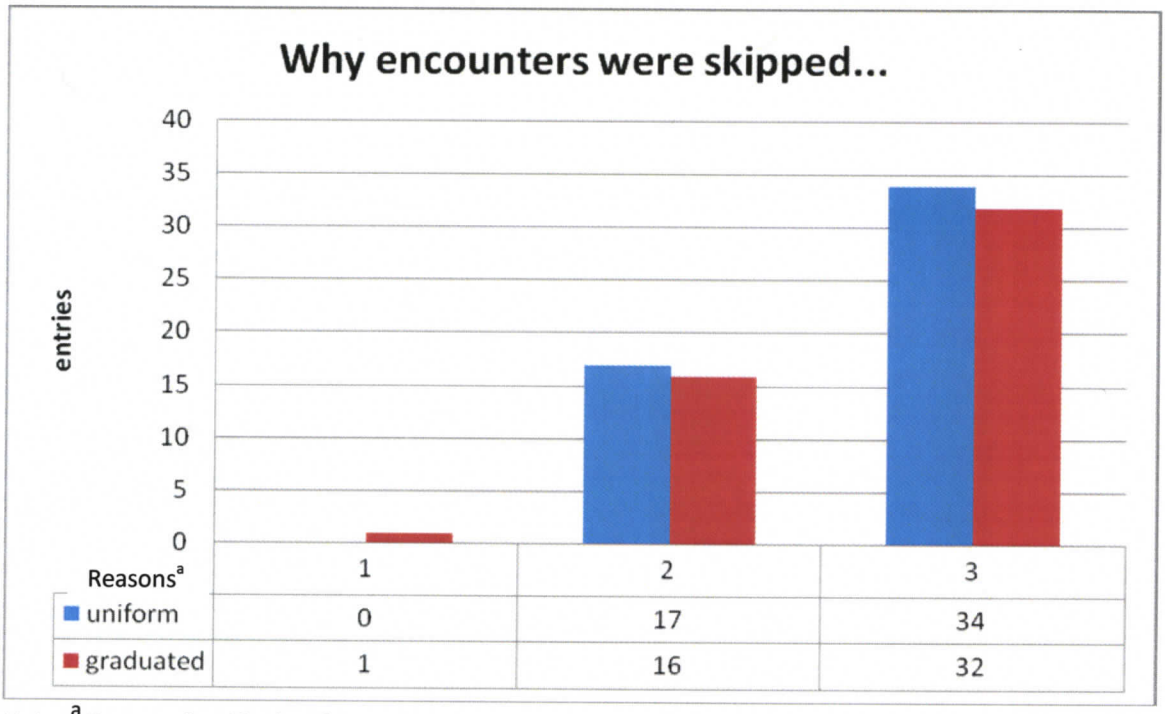


rating	1	2	3	4	5	6	7	8	9	10	11
uniform	0	1	11	11	8	6	3	0	4	0	0
graduated	1	8	7	11	3	5	2	3	0	0	0

Question 12

Question 12 requested students to report on reasons why they had skipped encounters. In particular, the first option, skipping a practice because they felt they did not need another one, was of interest for research question 2. Only one student chose this response. All others claimed that they had skipped because timing did not match their schedule (33 entries; UG = 17; GG = 16) or they had simply forgotten (66 entries; UG = 34; GG = 32).

Figure 18: Questionnaire B-12: Why were encounters skipped?



Note. [a] Reasons for skipping the practice sessions: 1= I did not need another one; 2= It did not fit my schedule; 3= I forgot

Question 13

Students' self- evaluation of their learning progress showed that they believed they needed at least four encounters in order to feel at ease with their learning progress.

The majority of 73 students (UG = 41; GG = 32) claimed they needed all five encounters (even though Question 12 demonstrated they did not always do them). Only 7 students (UG = 1; GG = 6) felt confident to skip encounters. But, they felt more comfortable skipping the second, third, and fourth encounter. For example, they argued:

Student 5: "*Studying in the beginning and then studying toward the end are the most important for me, the middle doesn't really matter.*"

Student 35: "*The first encounter teaches me vocab and the last refreshes it.*"

These typical student comments are reflected in the distribution analyses, too (see 4.3.1 above).

4.4 Summary

This section provided the research findings pertaining to validity issues, the first research question (interval schedules), and the second research question (practice frequency). It displayed evidence for this study's validity concerns on behalf of group homogeneity and the online research tool.

Next, it presented One-Way ANOVA analyses results for spaced interval schedules. There were no significant differences between uniform and graduated schedule even though higher mean averages seemed to favour the uniform schedule.

Finally, it used One-Way ANOVA analyses and qualitative analyses to compile frequency of encounter results. There was evidence for a significant advantage of 4/5 practice encounters over 2/3 encounters. Mostly, 4/5 encounters resulted in higher averages. Furthermore, students reported their need for five practice encounters. The following section will now discuss these findings.

5. Discussion of results and limitations

This section discusses the results in the same sequence as they were presented in the data analyses. First, I briefly address validity and argue for the homogeneous formation of UG and GG, the comparable settings, and for ViVo as an effective vocabulary learning tool. Then, I discuss the results the interval spacing and interval frequency studies yielded. Next, pedagogical implications are presented. Finally, I address limitations and further research needs.

5.1 Discussing validity

Validity concerns are particularly eminent in field research. This is not surprising as test conditions are much more difficult to control in this environment than in a laboratory setting. However, I argue that though SLVA research should continue to be inspired by laboratory findings in cognitive psychology it must transfer these results to the teaching environment of second language acquisition. Field experiments must be part of large scale research projects in spite of the limitations they face. The following sections illustrate how I have addressed these validity issues for this study.

5.1.1 Homogeneous groups

Section 3.3 provided information on behalf of students' background; sections 3.3.2 and 3.4.3 presented the filter criteria measures. These measures were used to create user profiles that ensured a homogeneous formation for UG and GG. Of the 117 students registered in four Ger100A sections, the data of only 86 were used. Due to the filter criteria measures, these students

- had had no prior exposure to the target language German
- had a similar social profile
- had a similar language heritage background

- had demonstrated a similar use of the learning tool ViVo
- had a similar learning background and preferences

Therefore, the test groups UG and GG could be regarded as closely homogeneous.

5.1.2 Research tool ViVo

This section discusses the validity of the research tool ViVo. In particular, it addresses the multimodal features and issues of student involvement. It then reports on the high test scores achieved when ViVo was used for practice.

Multimodal features

Section 3.4.4 described how and why ViVo's multimodal features were implemented. Recent research on the effects of multimedia supports the claim that web-based multimodal annotations improve vocabulary learning (Kim & Gilman, 2008; Rimrott, 2009). Furthermore, ViVo's multimodal features cater to different learner preferences. The outcome of this research is therefore not biased because one modality was unduly favoured. On the contrary, a closer look at the students' learner preferences expressed in their Questionnaire B portrays the diversity of the favourite features they enjoyed most. They mention spell-check, images, audio, sample sentences, and testing features. Students had a choice to follow their own best study practice.

Student involvement

Next, students felt involved and informed about their vocabulary task. Nakata (2008) claimed that sharing with students how a program is designed, and why its features will contribute to long-term retention, leads to a higher acceptance and motivation, eventually resulting in a higher learning gain (p. 17). The overall acceptance of ViVo as a device used in the beginner's course was high. This is also reflected in the fact that of 117 users only 8 were excluded from

the data analyses due to consistently poor participation (participation < 2 encounters).

High test scores

Finally, ViVo's effectiveness was examined. Both encounter groups (2/3 and 4/5) and both UG and GG achieved significantly higher scores on all print quiz parts studied with ViVo. Section 3.5 describes the statistical analyses employed, and section 4.1.2 presents the results that have led to this conclusion. Therefore, there can be little doubt that ViVo contributed significantly to the learning gain.

Conclusion

In conclusion, these results echo research findings on computer assisted language that express how CALL enriches the learning experience in general (Coady & Tozcu, 2004; Cobb, Horst, & Nicholae, 2005; Jones, 2004, 2006; Jones & Plass, 2002), how students benefit from multimedia glossing (Chun & Plass, 1996; Rimrott, 2009; Kim & Gilman, 2008; Yoshii, 2006), how vocabulary acquisition can be designed as an effective and enjoyable vocabulary learning experience (Loucky, 2005; Webb, 2007a, 2007b), and how online activities can enhance learners' focus on the learning task thus leading to better results.

In addition, the very fact that ViVo was a web-based tool allowed to control the test settings. Unlike a handheld device such as Leitner's *Lernkartei* (1972), ViVo controlled students' access to the material and tracked their performance. This advantage as a web-based research tool corresponds to other CALL research findings (Kim & Gilman, 2008; Nakata, 2008; Schmitt, 2008). Nakata (2008) and Mondria and Mondria-DeVries (1994) demonstrated the superiority of computer use over other vocabulary acquisition means such as

word lists or flashcards. Baturay et al. (2008) reached similar conclusions when testing their web-based product WEBVOCLE as efficient learning tool with spaced repetitions. It can therefore be concluded that the vocabulary online trainer ViVo was a valid research and learning tool for this study.

5.1.3 Comparable settings and test conditions

All participants were subjected to similar test conditions. Section 3.2 describes the setting, the learning material, the learner corpora, test corpora, and test procedures. They were the same for all participants. Even though the classes were taught by different instructors, this most likely did not lead to different teaching approaches because the instructors worked closely together. They planned and aligned their activities and assignments at weekly team meetings. Nevertheless, this is a limitation of a field research. Yet, I argue, even the same instructor does not respond in the same way in all of his or her classes. I therefore reason, that within the limits of a field study, the settings for test procedures and instruction compare.

5.2 Discussing interval schedule results

This research yielded no significant differences between uniform and graduated practice intervals. The following sections will look into two possible answers. First, differences do not exist. Second, they are not apparent within the setting of this research. This section then addresses possible limitations within the test setup.

5.2.1 Test results for uniform and graduated interval schedules

This research yielded no significant differences between the uniform practice schedule and the graduated practice schedule. Its findings on the first research question complement those of cognitive psychology research (i.e., Carpenter & DeLosch, 2005; Cull, 2000; Landauer & Bjork, 1978). They, too,

did not report overall significant evidence for the benefit of one interval spacing over the other. Advantages these researchers indicated were for example related to test types versus information repetitions (Landauer & Bjork, 1978) or attributed to delayed tests versus immediate recall (Karpicke & Roediger, 2007a). Moreover, most researchers examined interval encounters at the level of seconds, whereas this research contributed to our understanding of retention within a SLVA context: a typical chapter practice over a time period of two weeks. Cull's (2000) third and fourth test series provided a similar test setup to that of this study and his results correspond closely to the findings of this research. He, too, found that subjects performed equally well with uniform spacing and graduated spacing.

Karpicke and Roediger (2007b) have introduced the aspect that a uniform spaced practice schedule will produce better results because it delays initial retrieval and thus creates a task with a *desired difficulty*. They argued that an expanded schedule produced short-term benefits, but that a uniform spacing outperformed graduated spacing when tested with a 2-day delay. The findings of this study do not correspond with those of Karpicke and Roediger (2007b), even though uniformly practiced items did show slightly higher test results. Still, these test results obtained after a 4-day delay between last exposure and quiz did not show significant differences. Results for long-term retention – after weeks or even months – have yet to be compiled.

We can also conclude that the plateau of retention after a 2-day uniform interval was high enough to withstand forgetting with ease. This is supported by students' high test scores, their individual user log entries demonstrating only few recall errors, and students' self-evaluation of their learning effort between encounters (4.3.3). Pimsleur (1967) had claimed that a plateau of 60 % was the cross-over point for successful retrieval. Students in this research obtained higher

scores. If 60 % is indeed the threshold level, I argue that this would then indicate that the lengths between intervals could be expanded to more than two days. This is a condition that calls for further research.

Unlike Cull's (2000) research, this study was a field experiment with in-class SLV learning tasks. At the same time, the implementation of an online learning tool allowed for a closely monitored access and tracking of student data. In this, this research is unique and therefore contributes to our understanding of SLVA in language course settings. It has led to the following findings: Even though slightly higher test scores for the uniform condition were observed, these were not significant. This leads to two possible conclusions.

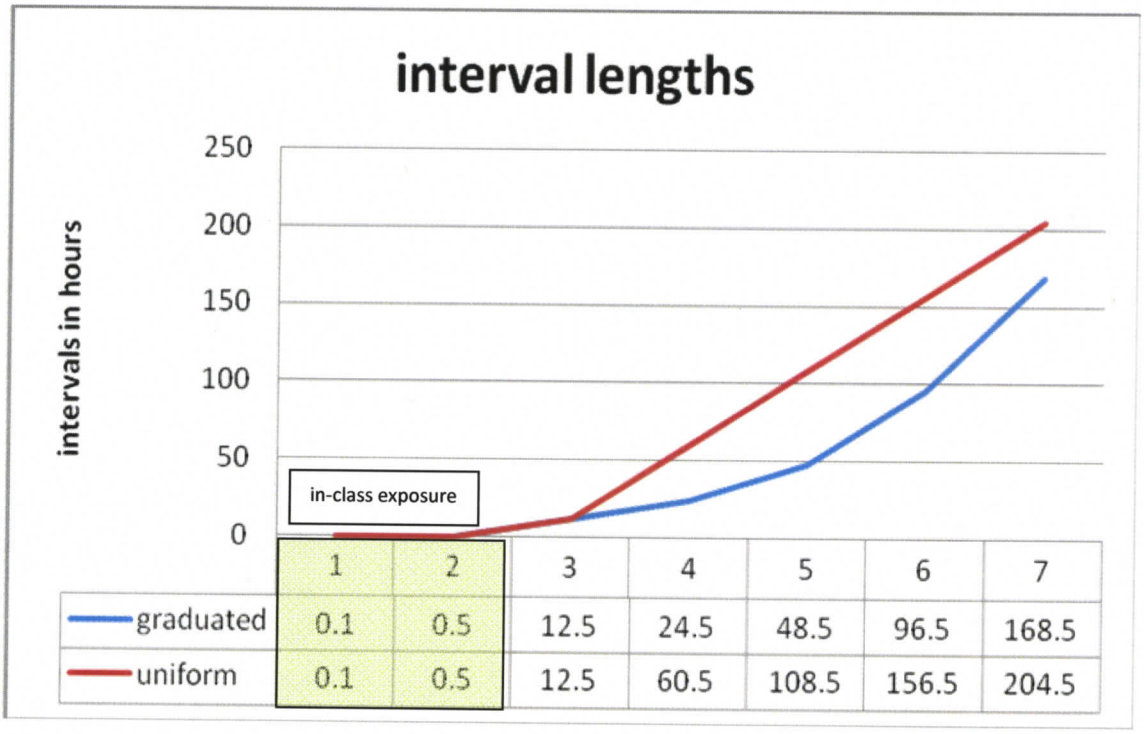
The first possibility unfolds as differences that simply do not exist as the above cited research findings imply. The second possibility reasons with more caution that these differences do not exist under the circumstances explored in this study. This second option considers that the data was compiled in a field experiment within the setting of post-secondary education and under the conditions of motivated L2 learners aged 20-25. Furthermore, the data refers to a beginner's class within an L1 environment and instruction in a European target language. These conditions led to limitations that are described in the following.

In-class exposure

One of these setting constraints is the fact that this research evaluated a learning process that was a segment of the learning curve. This learning curve already started after the first exposure of the target item in class and not with the first encounter on the learning platform ViVo. A closer look at the time schedule of this research scenario illustrates this point. First, instructors introduced ViVo words in class. They indicated their importance as target words for the vocabulary quizzes and presented them as part of the day's learning task. These first exposures were only roughly controlled by the research instructions to

discuss the learning task with ViVo and go over the vocabulary in class. These encounters were not timed nor recorded. Learning that may have occurred as result of these classroom activities was not part of the research setup. The table below indicates this phase (1 and 2) as part of the overall learning curve. Entering the vocabulary into the online trainer (ViVo encounter 1) on the day it was introduced in class was therefore the third exposure.

Figure 19: Interval lengths and in-class exposure



	1	2	3	4	5	6	7
graduated	0.1	0.5	12.5	24.5	48.5	96.5	168.5
uniform	0.1	0.5	12.5	60.5	108.5	156.5	204.5

Consequently, the statement on behalf of spacing intervals must include that there are no significant differences between uniform spaced and graduated intervals two weeks after entry into the spaced training sequences. In this, this research contributes to our understanding of SLV retention. Whereas cognitive psychology research mostly looked into short-term memory and working

memory this study focussed on retention over an intermediate time span within two weeks of first exposure.

Longitudinal research

The time issue leads to another limitation of this study's setup. *Robust learning* as defined by Van Lehn (2006) necessitates time. Even though retention was tested over a 2-week period, we cannot yet know how robust this achievement would prove to be over a much longer time span. This question needs to be addressed in quantitative longitudinal studies. One of the very few longitudinal studies involving spaced practice was Bahrick's et al. (1993). Their research was based on a 9-year qualitative observation of the four Bahrick family member's retention of a second language. However, they did not compare graduated with uniform intervals. They examined frequency issues and different interval lengths that were all uniform spaced. Therefore, further quantitative research comparing different interval spacing schedules is necessary.

Implications for future longitudinal research

The research setup designed for this thesis compiled data for the purpose of longitudinal research even though this data was not part of this thesis' data analyses. It consisted of a final print vocabulary quiz and an online end-of-term quiz. After 13 weeks of class instruction students wrote their final vocabulary quiz as part of their final exam. This exam was scheduled 2 weeks after the last day of class instruction. They also completed an end-of-term online vocabulary quiz[36]. The test corpora were composed of selected ViVo practiced and tested vocabulary items (4 for every chapter were chosen for the print quiz, 10 for every chapter were used for the online quiz).They adhered to the same format as all the other print respectively online quizzes. It follows that these test items were

[36] The data used for this thesis was compiled based on the 5 chapter print quizzes and the five chapter online quizzes with a maximum interval of 4 days between the chapter online quiz and the chapter print quiz.

subject to a long-term memory processing. Their retention times were staggered as follows

- chapter 1 items tested after 11 weeks
- chapter 2 items tested after 9 weeks
- chapter 3 items tested after 7 weeks[37]
- chapter 4 items tested after 4 weeks
- chapter 5 items tested after 2 weeks.

These tests will be part of a longitudinal study, but the fact that only 4 (print quiz), respectively 10 (online quiz) vocabulary items of every chapter were part of the final print and online quizzes led to an exclusion of this data in this study. I regarded a test corpus of 4 (10) target items per chapter as too small to elicit sufficient data for a quantitative research. For quantitative research purposes in future studies we must compile more test sets of these final quizzes. Next, I argue that additional follow-up tests after longer intervals should be scheduled.

5.2.2 Setup concerns and limitations

The section above discussed the interval schedule results. The following paragraph now addresses further objections and limitations. They are grouped as

- test setup concerns
- the human factor

Test setup concerns

In this section, I address concerns originating in the test set-up: in class vocabulary exposure, little differentiation due to high scores, and a time lapse between last encounter and tests.

---

[37] The Reading Break accounts for this change of interval.

Incidental encounters

One continuously researched aspect of SLVA is the influence of incidental encounters (i.e., Huckin & Coady, 1999; Hulstijn, 1992; Hulstijn & Laufer, 2001; Paribakht & Wesche, 1999; Watanabe, 1997; Webb, 2007b). Estimates on the number of exposures that are necessary to guarantee proficiency vary considerably. Zahar et al. (2001) stated that it would take 29 years to acquire a basic learner corpus, if incidental encounters were the only means to learn.

With this in mind, I reason that it is necessary to return to a more precise definition of the term *incidental* as discussed in section 2. Paribakht and Wesche (1999) defined that incidental learning occurs when students are focussed on comprehending the meaning but do not have the explicit goal of learning these words. I argue that, strictly speaking, the vocabulary items introduced in ViVo's learning cycle were not incidental encounters because they were items students were explicitly told they would be tested on. This view has been informed by Hulstijn and Laufer's (2001) definition: "Telling or not telling students that they will be tested afterwards on their knowledge is the critical operational feature distinguishing incidental from intentional (p. 267)". The students' exposure to the target items can only be characterized as *incidental* in the sense that students were not looking out for them but came across them unintentionally, albeit guided by the instructors' intervention. Therefore, I argue that the term *unintentional exposure* is more suitable.

Possibly, this exposure could have had a strong impact on student performance, and the frequency of these items could have then biased test results. One assumption was that students might have focused on these items because they knew that they would be part of their tests. Students encountered these test items in their readings, class instruction, and when completing assignments. However, it was surprising that findings of this research (see 4.1.1)

raised considerable doubt that this exposure had a strong influence on the students' learning gain.

The textbook frequency analysis showed that different items displayed a large scope in frequency (i.e., book– *das Buch*= 7; apartment–d*ie Wohnung* = 41). Yet, this frequency factor could not be related to the results of the quizzes. Students did not perform better on their tests for those items that had appeared more frequently in their textbook activities. A detailed error analysis of six randomly chosen and examined vocabulary items, as discussed in section 4.1.1, did not reveal performance differences. Those items that appeared less frequent in the textbook did not have lower scores on the print quizzes than those that appeared more frequently in the learning material. One possible explanation is that when students encountered these items their attention was not focused on learning the vocabulary but on understanding the communicative task or grammar point. The target items were therefore not acquired in these unintentional encounters. But, when studied with ViVo, a FonFs-type vocabulary training, students learned them. Any advantage that could have been related to a more frequent presentation of an item in the textbook was irrelevant when compared to the proficiency level acquired in the explicit vocabulary learning task. I reason, that the conscious effort put into this explicit vocabulary learning activity outweighed the learning gain of unintentional exposures. This view is in accordance with Laufer and Goldstein's (2004) studies which claim that depth of knowledge is closely related to the effort put into its acquisition. My claim is furthermore supported by the random sample comparison of the detailed user logs that documented what the students typed as answers and how often they asked for cues (see 4.1.1). These user logs did not yield an indication for different student behaviour. Both frequency conditions–more frequent and less frequent–did not only show the same error types of misspelling, gender mistakes, and

capitalization challenges, but also the same patterns of repeated trials. I therefore argue that the students did not feel more confident of their answer because they had encountered the items in the textbook or class.

I conclude that, based on the above presented arguments, the exposure to unintentional encounters most likely did not influence the learning outcome unduly. However, to explore this aspect in more depth, further evaluation of the present data with this particular focus in mind is necessary. This must then be accompanied by a research design that also observes and records unintentional exposure during class time instruction more closely.

Ceiling effect

A second aspect of test setup concerns are paradoxically the high scores that have created a ceiling effect. As much as instructors and students appreciated the excellent grades, these did not allow for a more differentiated distribution of results. With score averages ranging from 88.07 % to 95 % (letter grades of A and A+), basically everybody did well. Yet, based on my experience as an L2 instructor, the tests that were administered were by no means easier than those of previous years. Usually, students struggle with the acquisition process of vocabulary and the scores obtained in this study were therefore surprisingly high. Some might therefore argue that these results could be interpreted as an effect of *over-learning*. But on the other hand, this would then imply that less than 4/5 encounters were sufficient—a fact that is not undisputed as the discussion of the second research question has shown. I reason, that further research must address, if a different set-up with other intervals lengths and other frequencies would yield a *turning-point* where interval schedules would indeed matter. The results presented in this research can only be evaluated on the basis of this research's test set-up. Under these conditions, it holds that there were no significant differences between the uniform schedule and the graduated one. A

different research focus with differing interval lengths or frequency might still reveal these differences.

Time lapse between last practice and testing

The third test set-up issue concerns time lapses between last practice and testing. Due to their encounter schedule the graduated group finished their practice before the uniform did. This might have led to an advantage for the uniform group because the online test was closer in time for the UG (two days) than it was for the GG (four days). On the other hand, the interval between online quiz and print quiz was the same for everybody (4 days) and significant score differences between print quiz and online quiz did not occur. Because the ANOVA analyses (Table 22 and Table 23) revealed no significant score differences between these two dependant variables, it is unlikely that these time differences had a bearing on the results. By that time, the learning gain was already well established, so that online quizzes came up with the same result as the print quizzes and interval differences were not significant.

The human factor

Students' participation and positive attitude towards the learning tool has already been reported above. This section now takes a closer look at student test behaviour for the online quiz and the print quiz. Both quizzes had their pros and cons. The section then explores problems students faced when requested to complete the tasks according to schedule.

Online quiz

Each of the 5 online quizzes accounted for only 0.3 % of the final grade and was graded for completion. This fact may have led to the following student behaviour completing the online quizzes: Students skipped tests and students did

not aim to do well on the tests. The detailed data compiled in the Moodle user logs support this claim.

The Moodle user log shows a fluctuating participation that accounts for the first aspect. I term this the *achievement comfort zone*. Students skipped assignments because additional 0.3% would not affect their overall letter grade.

The second aspect was balanced by two facets: the possibility to test themselves or the last opportunity to skim through the material. The first scenario encompasses the following behaviour: Some students consider the online quiz to be a practice test quiz prior to the graded quiz. They therefore try to do well to self-test their proficiency level. This behaviour seemed to be the case especially for chapter 5 because this quiz was close to the final exam. The minimum scores were much higher (UG = 77 %; GG = 74 %) than in the other online quizzes. Furthermore, the Moodle protocols were much more detailed. They showed more error variations that documented students' repeated tries to get the correct answer. This indicated that students had tried to do as best as possible. Furthermore, compared to other chapters, they rarely skipped answers. In the second possible scenario, those students who regarded the online quiz as an extended practice opportunity did not mind the score because the online quiz was graded for completion only. Again, this behaviour could be observed in the Moodle user logs. Students completed the tests without regard to their test scores because they omitted to type in their responses. Yet, their test duration did not differ from other students who had typed in the complete answers. Furthermore, they had accessed all questions at least once, had requested clues and checked the answer. Their data had to be included in the ANOVA analyses because their tests were marked as completed.

In spite of these behavioural variants, using online quizzes in this research did have an advantage the print quizzes did not have. Most likely, students had

not studied outside the ViVo schedule to prepare for this quiz because they were under no pressure to do well. They could therefore test themselves without having to fear that low test results would affect their overall grades. This was a possible drawback for the print quiz.

Print quiz

One advantage for the print quiz data collection was the fact that it was graded, and I could justifiably assume students tried to do well. However, this very fact raised concerns that students would be tempted to add further practice sessions on their own. As mentioned before, I excluded students from the data analyses who had indicated this type of practice in their Questionnaire B (i.e., writing word lists daily, creating own flashcards, and using other online trainers). As another means to avoid students' wish to practice outside the assigned schedule, the online quiz was scheduled prior to the print quiz to counterbalance further practice needs. I reason, that if students used the online quiz to test themselves and they did well, they would then refrain from practicing more.

Finally, I argue that these factors have not biased this research because these conditions applied to both UG and GG. If after implementation of the screening criteria the test groups might still have had students who displayed this behaviour, chances are high that these would have been distributed evenly in both groups. Due to the large number of participants, these factors were therefore statistically not relevant.

Skipped encounters

Students skipped encounters. This aspect of observed student behaviour was the starting point for the second research question: Were more practice

encounters superfluous because robust retention could be achieved with fewer encounters?

It seemed contradictory that on the one hand students explicitly noted the necessity of 4/5 encounters yet neglected to complete them. In part, this could be attributed to the rigid time frame of uniform and, in particular, to the graduated schedule, but on the other hand, student behaviour observed in the pilot study 2008 showed that when given unlimited access many students did not make use of this option.

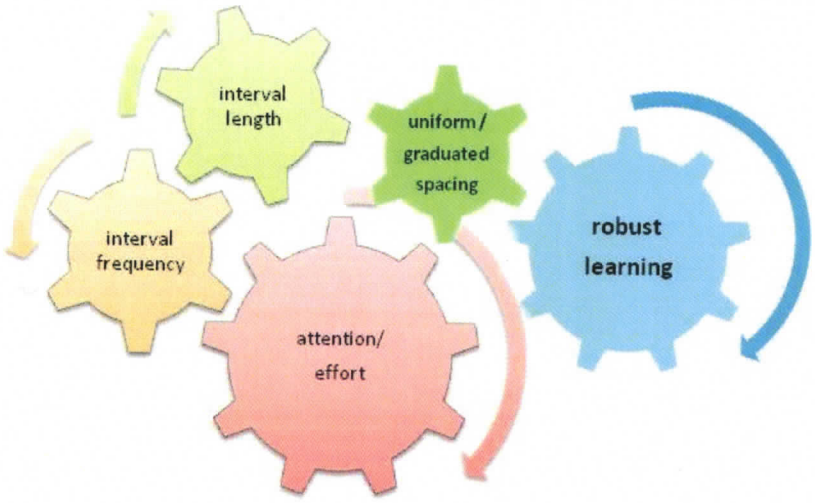
This phenomenon was also supported by data of a post-study completed in spring 2009. Two 100A sections were allowed 24/ 7 access to the vocabulary trainer and given the graded assignment to complete 5 encounters within the two-week chapter learning cycle. Preliminary results indicated that some students did not only disregard this option, they also accepted incomplete assignment grades. On the other hand, other students were overzealous and produced access rates of 190 %. This meant, they practiced almost twice as much as required. Still, we do not know if these practice sessions were necessary or not. Furthermore, we cannot know for sure, if students who claimed they needed 4/5 encounters were biased towards the myth of past pedagogical recommendations, the principle of *the more-the better,* or if they truly needed these encounters. They may have thought a frequency of 4/5 was necessary, whereas maybe fewer practices could have been sufficient. On the other hand, the print quiz scores were indeed significantly higher when items had been practiced 4 or 5 times. I therefore conclude that most students would very likely benefit significantly from 4 or 5 encounters without *over-learning.* But, this study does not provide sufficient data to mark a possible turning point where effort input is balanced by the learning factor output. Again, this needs to be addressed in further research.

Furthermore, I would also like to point out other influences on the learning gain this research could not account for. Further research is necessary to elicit data on behalf of students' motivation, students' observed behaviour, and students' self-evaluation. These data sets must then be correlated to quiz test scores because this data too is linked to the spacing of practices. For example, how would students perform, if given the choice of randomly spaced intervals that they control? Would this option of student-centered access take precedent over other factors such as spacing and frequency? Next, this frequency discussion is explicitly linked to a FonFs-type learning task. Different learning objectives with regard to proficiency level; receptive or productive skills; and based on speaking, listening, writing, and reading tasks will most likely differ in their frequency requirement. This too, needs to be explored in further research.

5.3. Pedagogical implications

When striving for an optimal balance between time investment, effort, manageability, and learning outcome, we come across the factors of interval frequency, interval length, and spacing schedules. Each one of them has an impact on *robust* learning and retention. However, as much as we need to explore them individually, we also need to see them as interrelated processes. For example, will shorter interval lengths determine how often a learner needs to study? Most likely, the effort, motivation, and attention a learner puts forward to the learning task will affect the interval frequency. Further research needs to examine these relationships. These factors will shift depending on the depth of knowledge and the proficiency level in the subskills reading, writing, listening, and speaking the learner wishes to achieve.

Figure 20: Pedagogical implications- interrelated factors



As mentioned before, the findings of the first research question concur with Cull's research (2000). They can therefore contribute to our understanding of SLVA because they might change how we view recommendations on spaced learning in educational settings. Both Cull's and this study contradict and challenge what pedagogical practice has embraced and taken for granted for many years (i.e., Oxford, 1990, 2003; Pimsleur, 1967). Most second language acquisition textbooks still recommend and favour a graduated schedule (i.e., Folse, 2004; Kennedy, 2006; Nation, 2001; Oxford, 1990; Schmitt, 2002).

However, best practice learning tasks balance feasibility and accomplishment. With this in mind, I favour a uniform interval schedule. Though the graduated schedule and the uniform schedule did not differ significantly, the uniform average outcomes were slightly higher. But more importantly, the task was more manageable. It was easier for students to remember they needed to study every other day. The GG struggled more with time issues. Next, I argue for

this study pattern, because the plateau of remembering was evidently still high enough to guarantee recall with ease. This is reflected in the students' rating of difficulty recalling the learning task between the encounters (Questionnaire B). Even though the GG reported this to be easier, both groups did not claim problems remembering.

With regard to frequency of encounters, 4 or 5 encounters for this type of FonFs task seemed best practice and significant evidence points into this direction for online quizzes and print quizzes. Students felt more confident of their learning gain. Yet, the obtained scores for 2/3 and 4/5 encounters may have created a ceiling effect that does not allow for a more precise differentiation of results. Therefore, further research is necessary. This research could then also explore how the number of necessary practice sessions depended on individual learner preferences and differing learner abilities. I argue that– in spite of the many variables– a research based curriculum design should aim to provide guidelines for the question which practice frequency for the various tasks and proficiency levels would suit most students.

5.4 Conclusion

Students practicing on a uniform practice schedule and those working on a graduated schedule performed equally well. The findings for the first research question did not elicit significant differences between a graduated interval schedule and a uniform interval schedule. Neither online quizzes nor print quizzes differed in this result. However, UG mean scores for both test procedures were slightly higher. These results complement and correspond to previous research, but they are not in agreement with pedagogical practice that favours a graduated interval schedule. It is therefore advisable that future curriculum design views spaced practice schedules more critically. At the same time, I have pointed out the need for further research with regard to aspects that could not

be covered within the context of this thesis. It also holds that SLVA must be viewed within the larger context of other disciplines. I therefore conclude with a brief outline of future perspectives in the section below.

6. SLVA research – future perspectives

Learners, instructors, and researchers agree, that vocabulary acquisition is among the most daunting tasks language learners face. Instructors, textbook authors, and curriculum designers need research insights on how to provide learning environments that are informed by pedagogically good choices. This section describes future research needs and shows how the research presented in this study contributes to our understanding of spaced learning. But at the same time, it points out the need for future interdisciplinary research. It concludes with the imagery of a *research cycle* conveying the notion that no research comes full circle that does not spark future research interests.

6.1 Interdisciplinary approach

In the wake of new emerging technologies allowing for a closer glimpse at governing principles of language processing, knowledge of all disciplines involved must be combined towards a comprehensive understanding of the structures and processes involved in language learning. In the past years, neurology has informed many research issues in cognitive psychology and linguistics. Yet, there still seems to be a gap between disciplines. In 1992, Jacobs and Schumann suggest that "language acquisition researchers must begin to incorporate a degree of neurobiological reality into their perception of the language acquisition process" (Jacobs & Schumann, 1992, p. 282). Fifteen years later, Hulstijn (2007) still cautions against the divide in the research pertaining to neurophysiology and linguistics. In order to truly understand the processes underlying the fundamental mechanisms of language acquisition, he recommends that researchers of all disciplines "have to work together on the mind-side of the mind/brain coin" (Hulstijn, 2007, p. 15).

This view is shared by educators, linguists, and cognitive psychologists alike (i.e., Aitchison, 2003; Hulstijn, 2007; Scherfer, 1994). Neurologists, too, call for interdisciplinary research in order to understand language acquisition processes in their complexity (Friederici et al., 2002). In consequence, SLVA research must do both: it must rely on findings of these disciplines, but also contribute to research within their own field.

This study has presented neurological research on processes and structures that govern language in our brain, and it has addressed cognitive psychology research pertaining to memory. This was followed by research findings on behalf of the mental lexicon, and finally, this study focused on linguistic theories of the mental lexicon with regard to SLVA.

Based on this theoretical background, this thesis then presented the research design on behalf of vocabulary retention focusing on the aspect of a decontextualized cyclical practice. In particular, this research has thereby contributed to our understanding of the importance implementing spaced rehearsal sessions. At the same time, it has marked the preference of a graduated spaced study over a uniform spaced practice as controversial and has provided evidence that one practice interval schedule should not take precedent over the other. I have therefore advocated for more caution when implementing these schedules in SLV curriculum design. At the same time, I have pointed out the need for further research, in particular, exploring the interrelatedness of the factors frequency, interval schedules, task design, interval lengths, and robust learning gain.

6.2 Research cycle

Seliger and Shohamy (1989) have coined the term *research cycle*. Every research finding opens the door to yet further questions. This research contributed to the question, if students would benefit more from a graduated or

from a uniform spaced vocabulary practice schedule. Further research could diversify these findings. For example, it could explore if different word categories account for different spacing needs, or if a combination of various presentation modes work best with certain spacing schedules.

Most prominent, however, is the research interest concerning long-term retention. As discussed above, this research was conducted within the time-frame of a 13-week university term. This time frame needs to be expanded to introduce a longitudinal study. Both Read (2004) and Nation (2001) described vocabulary acquisition with all its aspects as a procedural continuum (2.1.1) This research could only partially contribute to this notion of continuity because the features of the tool, which was used in this research, could only partly cater to this concept of procedural acquisition: It had time constraints and design limitations. Possibly, future design variants of this tool, which for example could expand as well as repeat lexical entries, would then contribute further to our understanding of retention and spaced intervals.

Next, this study was conducted in the environment of postsecondary education. Its results may not transfer to secondary or primary education without reserve. This calls for further research to explore correlations of age, gender, and educational background on spaced practice schedules.

Finally, real time in a real setting for this study was in itself a limitation. For example, it limited the learner corpora to that of the textbook used. A research design within a more controlled environment would allow for more controlled tasks (i.e., examining cognates, different word categories, or annotation variants.

Last, this was a quantitative research. A mixed research that includes qualitative research might address issues of a correlation between motivation, learning styles, perceived proficiency level, and retention in a CALL environment.

Bibliography

Aitchison, J. (2003). *Words in the mind: An introduction to the mental lexicon.* Oxford, UK: Blackwell Pub.

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

Arbinger, R. (1984). *Gedächtnis*.[Memory] Darmstadt, Germany: Wissenschaftliche Buchgesellschaft.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation* (pp. 89–195). New York: Academic Press.

Baddeley, A. D. (1986). *Working memory*. Oxford, UK: Clarendon.

Baddeley, A. D. (1990). *Human memory*. London: Lawrence Erlbaum Associates.

Baddeley, A. D. (1997). *Human memory: Theory and practice*. Hove, UK: Psychology Press.

Baddeley, A. D. (2007). *Working memory, thought, and action*. New York: Oxford University Press.

Baddeley, A. D., & Hitch, G. (1974). Working memory. In G.A.Bower (Ed.), *Vol. 8. Recent advances in learning and motivation* (pp. 47-89). New York: Academic Press.

Baddeley, A.D., & Wilson, B. (1994). Errorless learning in the rehabilitation of memory-impaired people. *Neurophysiological Rehabilitation,* 4(3), 307-326.

Bahrick, P., Bahrick, H., Bahrick, L., & Bahrick, A. (1993). Maintenance of foreign language vocabulary and the spacing effect. *Psychological Science, 4*(5), 316-321.

Bahrick, H., & Phelphs, E. (1987). Retention of Spanish vocabulary over 8 years. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*(2), 344-349.

Balota, D., Duchek, J., & Logan, J. (2007). Is expanded retrieval practice a superior form of spaced retrieval? A critical review of the extant literature. In J. Nairne (Ed.), *The foundations of remembering* (pp. 83-105). London: Psychology Press.

Başar, E. (2004). *Memory and brain dynamics: Oscillations integrating attention, perception, learning, and memory*. London: CRC Press.

Baturay, M., Yıldırım, S., & Daloğlu, A. (2008). Effects of web-based spaced repetition on vocabulary retention of foreign language learners. *Egitim Arastirmalari Eurasian Journal of Educational Research,* 34, 1-21.

Bjork, R. (1994) Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185-205). Cambridge: MIT Press.

Bloom, K. C., & Shuell, T. J. (1981). Effects of massed and distributed practice on learning and retention of second language vocabulary. *Journal of Educational Research, 74*, 245-248.

Carpenter, S. K., & DeLosch, E. L. (2005). Application of the testing and spacing effects to name learning. *Applied Cognitive Psychology, 19*, 619-636.

Carroll, S. E. (1992). On cognates. *Second Language Research, 8*(2), 93-119.

Chun, D. M., & Plass, J. L. (1996). Effects of multimedia annotations on vocabulary acquisition. *Modern Language Journal, 80*(2), 183-198.

Coady, J., & Tozcu, A. (2004). Successful learning of frequent vocabulary through CALL also benefits reading comprehension and speed. *Computer Assisted Language Learning, 17*(5), 473-495.

Cobb, T., Horst, M., & Nicholae, I. (2005). Expanding academic vocabulary with an interactive on-line database. *Language Learning and Technology, 9*(2), 90-110.

Cohen, A. (2003). The learner's side of foreign language learning: Where do styles, strategies, and tasks meet? *International Review of Applied Linguistics in Language Teaching, 41*(4), 279-291.

Conway, A., Jarrold, C., Kane, M., Miyake, A., & Towse, J. (2007). *Variation in Working Memory*. New York: Oxford University Press.

Cowan, N. (1996). Short-term memory, working memory, and their importance in language processing. *Topics in Language Disorders, 17*, 1-18.

Cowan, N. (2001). The magical number 4 in short- term memory: A reconsideration of mental storage capacity. *Behaviour and Brain Sciences,* (24), 87-114.

Cowan, N. (2005). *Working memory capacity*. Hove, UK: Psychology Press.

Craik, F., & Lockhart, R. S. (1972). Levels of processing. A framework for memory research. *Journal of Verbal Learning and Verbal Behaviour, 11*, 671-684.

Crystal, D. (1987). *The Cambridge encyclopaedia of language*. Cambridge: Cambridge University Press.

Cull, W. (2000). Untangling the benefits of multiple study opportunities and repeated testing for cued recall. *Applied Cognitive Psychology, 14*(3), 215-235.

Daloğlu, A., Baturay, M., & Yıldırım., S. (2008). Designing a constructivist vocabulary learning material. In V. de Cassia, R. Marriott & P. Torres (Eds.), *Handbook of research on E-learning methodologies for language acquisition* (pp. 186-203). Hershey, PA: Idea.

De la Fuente, M. J. (2002). Negogiation and oral acquisition of L2 vocabulary: The role of input and output in the receptive and productive acquisition of words. *Studies in Second Language Acquisition,* (24), 81-112.

DeKeyser, R. M. (1998). Beyond focus on form: Cognitive perspective on learning and practical second language grammar. In C. Doughty & J. Williams (Eds.), *Focus on form in classroom second language acquisition* (pp. 42-63). Cambridge: Cambridge University Press.

Dempster, F. N. (1987). Effects of variable encoding and spaced presentation on vocabulary learning. *Journal of Educational Psychology, 79*, 162-170.

Di Donato, R., Clyde, M. D., & Vansant, J. (2007). *Deutsch Naklar!* (5th ed.). New York: McGraw.

Doughty, C. (2001). Cognitive underpinnings of focus on form. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 206-257). Cambridge: Cambridge University Press.

Ebbinghaus, H. (1885). *Über das Gedächtnis: Untersuchungen zur experimentellen Psychologie.*[About memory: research in experimental psychology] Leipzig, Germany: Duncker und Humblot.

Ebbinghaus, H. (1913). *Grundzüge der Psychologie*. [Fundamentals of psychology] Leipzig, Germany: Viet.

Ebbinghaus, H. (1964). *Memory: A contribution to experimental psychology*. New York: Dover Publications.

Ellis, N. (2003). Memory for language. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 33-68). Cambridge: Cambridge University Press.

Ellis, N., & Beaton, A. (1993). Factors affecting the learning of foreign language vocabulary: Imagery keyword mediators and phonological short-term memory. *The Quarterly Journal of Experimental Psychology, 46* (3), 533-558.

Ellis, R. (1995). Modified oral input and the acquisition of word meanings. *Applied Linguistics, 16*(4), 409-444.

Ellis, R., Basturkmen, H., & Loewen, S. (2002). Doing focus-on-form. *System,* (30), 419-432.

Ellis, R., & Xien, H. (1999). The roles of modified input and output in the incidental acquisition of word meanings. *Studies in Second Language Acquisition, 21*, 285-301.

Folse, K. S. (2004). *Vocabulary myths: Applying second language research to classroom teaching*. Ann Arbor: University of Michigan Press.

Freebody, P., & Anderson, R. C. (1983). Effects of vocabulary difficulty, text cohesion, and schema availability on reading comprehension. *Reading Research Quarterly, 18*(3), 277-294.

Friederici, A.D. (1985). Levels of processing and vocabulary types: evidence from on-line comprehension in normals and agrammatics. *Cognition,* 19, 133-166.

Friederici, A. D. (2000). The developmental cognitive neuroscience of language: A new research domain. *Brain and Language, 71*(1), 65-68.

Friederici, A. D. (2002). Wie wir Sprache verstehen - neuronale Präzision in Raum und Zeit. How we conceive language- neural precision in space and time] *Max-Planck Institut für Neuropsychologische Forschung, Jahrbuch 2002,* 43-53. Retrieved May 11, 2009, from http://www.mpg.de/pdf/jahrbuch_2002/jahrbuch2002_043_054.pdf

Götze, L. (1997). Hirnprozesse und die Rolle des Gedächtnisses beim Lesen.[Processes of the brain and the role of memory in reading] *Materialien Deutsch als Fremdsprache.* (46), 85-94.

Götze, L. (1997). Was leistet das Gehirn beim Fremdsprachenlernen.[What does the brain accomplish learning a foreign language] *Zeitschrift für Interkulturellen Fremdsprachenunterricht, 2*(2), 1-15.

Groot, P. J. M. (2000). Computer assisted second language vocabulary acquisition. *Language Learning & Technology, 4*(1), 60-81.

Horst, M., & Cobb, T. (2006). Second language vocabulary acquisition. *The Canadian Modern Language Review, 63 (1), 1-12.*

Huckin,T., & Coady, J. (1999). Incidental vocabulary acquisition in a second language. *Studies in Second Language Acquisition, 21* , 181-193.

Hulstijn, J. (1992). Retention of inferred and given word meanings: Experiments in incidental vocabulary learning. In P. J. Arnaud, & H. Bejoint (Eds.), *Vocabulary and applied linguistics* (pp. 113-125). London: MacMillan.

Hulstijn, J. (1997). Mnemonic methods in foreign language vocabulary learning: Theoretical considerations and pedagogical implications. In J. Coady & T. Huckin (Eds.), *Second language vocabulary acquisition* (pp. 203-224). Cambridge: Cambridge University Press.

Hulstijn, J. (2002). What does the impact of frequency tell us about the language acquisition device? *Studies in Second Language Acquisition,* (24), 269-273.

Hulstijn, J. (2003). Intentional and incidental second language vocabulary learning: A reappraisal of elaboration, rehearsal and automaticity. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 258-286). Cambridge: Cambridge University Press.

Hulstijn, J. (2007). Fundamental issues in the study of second language acquisition. *EuroSLA Yearbook 2007,* Retrieved May 11, 2009, from http://home.medewerker.uva.nl/j.h.hulstijn/bestanden/Hulstijn_Fundamental_issues_EuroSLA_Yearbook_2007.pdf

Hulstijn, J., & Laufer, B. (2001). Some empirical evidence for the involvement load hypothesis in vocabulary acquisition. *Language Learning, 51*(3), 539558.

Jacobs, B., & Schumann, J. (1992). Language acquisition and neurosciences: Towards a more integrative perspective. *Applied Linguistics, 13*(3), 282301.

Jones, L. C. (2004). Testing L2 vocabulary recognition and recall using pictorial and written test items. *Language Learning and Technology, 8*(3), 122143.

Jones, L. C. (2006). Effects of collaboration and multimedia annotations on vocabulary learning and listening comprehension. *CALICO Journal, 24*(1) 3358.

Jones, L. C., & Plass, J. L. (2002). Supporting listening comprehension and vocabulary acquisition in French with multimedia annotations. *Modern Language Journal, 86*(4), 546561.

Jones, R. L., & Tschirner, E. (2006). *A frequency dictionary of German*. London: Routledge.

Kandel, E. (2007). *In search of memory - The emergence of a new science of mind.* New York: Norton & Company.

Karpicke, J. D., & Roediger, H. L. (2007a). Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 33*(4), 704719.

Karpicke, J., & Roediger, H. (2007b). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language, 57*(2), 151162.

Kennedy, T. (2006). *Brain-based learning through content- based language teaching*. Retrieved May 11, 2009 from http://www.flbrain.org/research.htm

Kim, D., & Gilman, D. (2008). Effects of text, audio, and graphic aids in multimedia instruction for vocabulary learning. *Educational Technology and Society, 11*(3), 114126.

Landauer, T., & Bjork, R. (1978). Optimal rehearsal patterns and name learning. In M. Grüneberg, P. Morris & R. Sykes (Eds.), *Practical aspects of memory* (pp. 625632). London: Academic Press.

Laufer, B. (2006). Comparing focus on form and focus on FormS in second language vocabulary learning. *Canadian Modern Language Review/ La Revue Canadienne Des Langues Vivantes, 63*(1), 149166.

Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning, 54*(3), 399436.

Leitner, S. (1972). *So lernt man Lernen: Der Weg zum Erfolg*.[How to learn to learn: The road to success] Freiburg, Germany: Herder.

Levelt, W. (1993). Language use in normal speakers and its disorders. In G. Blanken, H. Dittmann, H. Grimm, J. Marshall & C. Wallesch (Eds.), *Linguistic disorders and pathologies: An international handbook* (pp. 115). Berlin: de Gruyter.

Loucky, J. P. (2005). Combining the benefits of electronic and online dictionaries with CALL web sites to produce effective and enjoyable vocabulary and language learning lessons. *Computer Assisted Language Learning, 18*(5), 389-416.

Lüders, J. (2005). Wortschatzlernen mit der Lernkartei? Anspruch und Wirklichkeit.[Vocabulary learning with the flashcard system? Allegations and reality] *PRAXIS Fremdsprachenunterricht, 1*, 24-27.

Marslen-Wilson, W.D. (1987). Functional parallelism in spoken word recognition. *Cognition, 25*, 71-102.

Marslen-Wilson, W.D., & Welsh, A. (1978). Processing interactions and lexical access during word-recognition in continuous speech. *Cognitive Psychology, 10*, 29-63.

Max-Planck-Institut Pressemitteilungen. (2006) Wie man ‚Kontakt' hält[How to stay in 'contact']. Retrieved June 10, 2009 from http://www.mpg.de/bilderBerichteDokumente/dokumentation/pressemitteilungen/2006/pressemitteilung200601121/index.html

Mondria, J. (2003). The effects of inferring, verifying, and memorizing on the retention of L2 word meanings. *Studies in Second Language Acquisition, 25*, 473-499.

Mondria, J., & Mondria-De Vries, S. (1994). Efficiently memorizing words with the help of word cards and "hand computer": Theory and applications. *System, 22*(1), 47-57.

Mueller, J. L., Rüschemeyer, S., & Friederici, A. D., S. (2006). Aktivitätsmuster im Gehirn: Unterschiede und Gemeinsamkeiten beim Verstehen von Erst- und Zweitsprache [Mapping neural activity in the brain: different and identical aspects understanding first and second language]. *Neuro Forum: Perspektiven der Hirnforschung,* (6), 176-184.

Nakata, T. (2008). English vocabulary learning with word lists, word cards and computers: Implications from cognitive psychology research for optimal spaced learning. *ReCALL, 20*(1), 3-20.

Nassaji, H. (2004). The relationship between depth of vocabulary knowledge and L2 learners' lexical inferencing strategy, use and success. *The Canadian Modern Language Review, 61*(1), 107-134.

Nation, I. S. P. (1990). *Teaching and learning vocabulary*. Rowley, MA: Newbury House.

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

Oxford, R. L. (1990). *Language learning strategies: What every teacher should know*. Boston: Heinle & Heinle.

Oxford, R. L. (2003). Language learning styles and strategies: Concepts and relationships. *International Review of Applied Linguistics in Language Teaching, 41*, 271-278.

Paradis, M. (1987). *The assessment of bilingual aphasia*. Hillsdale, NJ: Erlbaum.

Paribakht, T. S., & Wesche, M. (1999). Reading and "incidental" L2 vocabulary acquisition. *Studies in Second Language Acquisition, 21*, 195-224.

Pimsleur, P. (1967). A memory schedule. *The Modern Language Journal, 51*(2), 73-75.

Raupach, M. (1994). Das mehrsprachige mentale Lexikon.[The multilingual mental lexicon] In W. Borner & K. Vogel (Eds.), *Kognitive Linguistik und Fremdsprachenerwerb* (pp. 19-37). Tübingen, Germany: Narr.

Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language testing10(3)*, 355-371.

Read, J. (2004). Plumbing the depths: How should the construct of vocabulary knowledge be defined? In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language: Selection, acquisition, and testing.* (pp. 209-227). Amsterdam: John Benjamins.

Rimrott, A. (2009, March). Adaptive vocabulary instruction for L2 learners of German. Paper presented at the CALICO conference, Arizona State University.

Risager, K. (2006). *Language and culture*. Clevedon, UK : Multilingual Matters.

Roediger, H., L., & Karpicke, J. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1*(3), 181-210.

Rott, S. (1999). The effect of exposure frequency on intermediate language learner's incidental vocabulary acquisition through reading. *Studies in Second Language Acquisition, 21*(1), 589-619.

Scherfer, P. (1994). Überlegungen zu einer Theorie des Vokabellernens und -lehrens.[Considerations on behalf of a vocabulary learning and teaching theory] In W. Berner & K. Vogel (Eds.), *Kognitive Linguistik und Fremdsprachenerwerb* (pp. 185-215). Tübingen, Germany: Gunter Narr.

Schmidt, R. (1994). Deconstructing consciousness: In search of useful definitions for applied linguistics. *AILA Review, 11,* 11-26.

Schmidt, R. (2001). Attention. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 3-32). Cambridge: Cambridge University Press.

Schmitt, N., & Schmitt, D. (1995). Vocabulary notebooks: Theoretical underpinnings and practical suggestions. *ELT Journal, 49*(2), 133-143.

Schmitt, N. (2000). *Vocabulary in language teaching*. New York: Cambridge University Press.

Schmitt, N. (2002). *An introduction to applied linguistics*. New York: Oxford University Press.

Schmitt, N. (2008). Teaching vocabulary. Longman teacher guidelines series. Retrieved May 11, 2009, from http://www.nottingham.ac.uk/english/lookup/lookup_az.php?id=NjAxNjc0&page_var=personal

Schröder, B., & Roedig, T.(n.d.) *Vermittlung von Lernstrategien: Skript für die Beratungslehrer-Ausbildung im Regierungsbezirk Karlsruhe.* [Teaching learner strategies- teacher training handout for the school district Karslruhe]. Retrieved May 11, 2009, from http://www.schule-bw.de/lehrkraefte/beratung/beratungslehrer/probleme/lat/lernstrategien.pdf

Seibert, L.C. (1927). An experiment in learning French vocabulary. *Journal of Educational Psychology, 18*(5), 294-309.

Seliger, H. W., & Shohamy, E. (1989). *Second language research methods*. New York: Oxford University Press.

Singleton, D. (1999). *Exploring the second language mental lexicon*. Cambridge: Cambridge University Press.

Singleton, D. (2005). The critical period hypothesis: A coat of many colours. *International Review of Applied Linguistics in Language Teaching, 43*(4), 269-285.

Smith, B. (2004). Computer-mediated negotiated interaction and lexical acquisition. *Studies in Second Language Acquisition, 26*, 365-398.

Tombaugh, T., & Hubley, A. (2001). Rates of forgetting on three measures of verbal learning using retention intervals ranging from 20 min to 62 days. *Journal of International Neuropsychological Society, 7*, 79-91.

Van Lehn, K. (2006). *The PSLC theoretical framework.* Retrieved May 11, 2009, from http://learnlab.org/clusters/PSLC_Theory_Frame_June_15_2006.pdf

Watanabe, Y. (1997). Input, intake, and retention: Effects of increased processing on incidental learning of foreign language vocabulary. *Studies in Second Language Acquisition, 19*(3), 287-307.

Webb, S. (2007a). The effects of repetition on vocabulary knowledge. *Applied Linguistics, 28*(1), 46-65.

Webb, S. (2007b). Learning word pairs and glossed sentences: The effects of a single context on vocabulary knowledge. *Language Teaching Research, 11*(1), 63-81.

Wilson, B., Baddeley, A., Evans, J., & Shiel, A. (1994). Errorless learning in the rehabilitation of memory impaired people. *Neuropsychological Rehabilitation, 4*(3), 307-326.

Yoshii, M. (2006). L1 and L2 glosses: Their effects on incidental vocabulary learning. *Language Learning and Technology, 10*(3), 85-101.

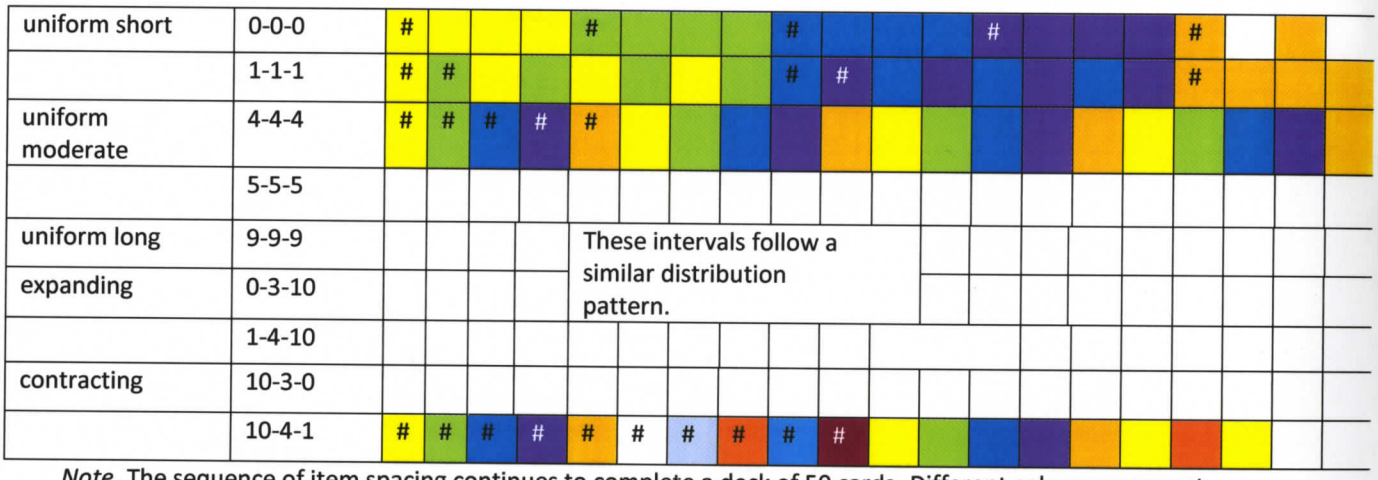
Zahar, R., Cobb, T., & Spada, N. (2001). Acquiring vocabulary through reading: Effects of frequency and contextual richness. *The Canadian Modern Language Review, 57*(4), 541-572.

Appendix

Appendix A

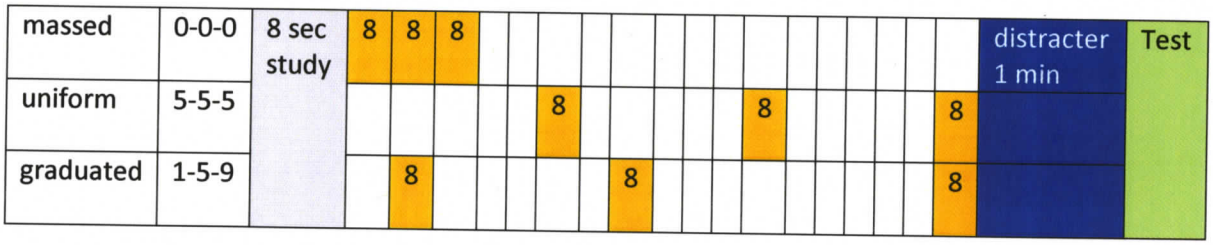
Table A1: Interval patterns for Landauer and Bjork's test series

The colour coding of the table reflects the sequential arrangement within the 50 cards of the deck and is set up for illustration purposes. Every deck contained two different interval schedules. Depending on the arrangement of the cards these then formed a spacing pattern. All items were presented for 9 seconds each. Twelve target items (first name–last name pairs) were studied and then presented as test (first name only with a blank to fill in the last name).



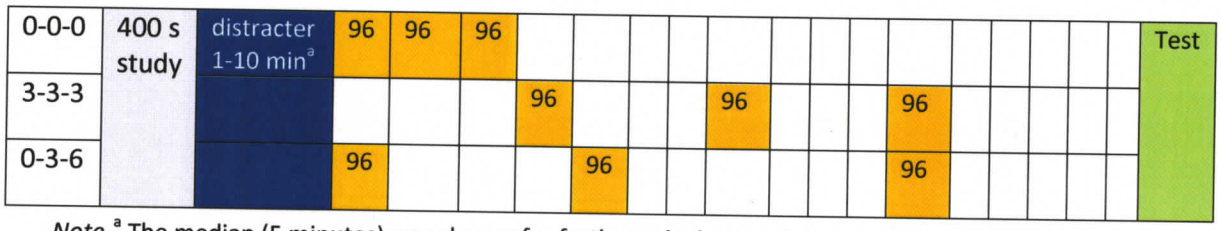
Note. The sequence of item spacing continues to complete a deck of 50 cards. Different colours represent different target items. The # denotes the first study encounter of these items. This experiment used 4 filler items as buffer.

Table A2: Interval patterns for Cull's first test series



An analysis of Cull's (2000) spacing patterns enabled the calculation of cumulative time lapses between first encounter and further encounters measured in seconds. Items were presented for 8 seconds each. This rehearsal period was followed by a one minute distracter. The 32 target items were then tested in a cued recall test. Uniform and graduated condition had their last encounter after 144 seconds; all conditions were tested after 212 seconds.

Table A3: Interval patterns for Cull's second test series



Note.[a] The median (5 minutes) was chosen for further calculation of the tables.

In his second series of experiments, Cull provided a study phase of 400 seconds for his 8 target items. Subjects worked at their own pace, multiple times, and with spaced intervals. The rehearsal patterns were then administered after a 1-10 minute distracter. The 8 target items were presented 12 seconds each (total 96 s). Cull then created his practice schedules by spacing them with distracter intervals (2-3 minute long ratings of word knowledge). His practice schedules were massed 0-0-0, uniform spaced 3-3-3, and expanded 0-3-6.

Table A4: Interval patterns for Cull's third and fourth test series

0-0-0[a]	30 min study											3 days (series 3)	TEST
0-0-0												or	
1-2-3		3 										9 days (series 4)	
2-2-2		3 											

Note.[a] For Cull's third and fourth test series, the intervals are measured in days.

 This icon represents the massed rehearsal of all three test booklets.

 This icon represents the rehearsal of one test booklet.

 This icon represents the distribution of filler tasks.

Table A5: Interval patterns for Karpicke and Roediger's first test series

Karpicke & Roediger (2007b) – Experiment 1

The tables below show Karpicke and Roediger's test setup and test conditions.

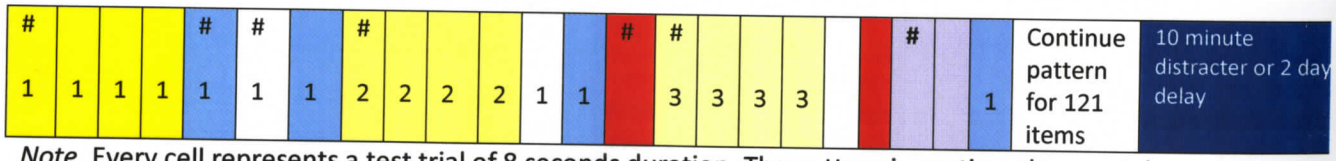
	# 0-0-0	# 1-5-9	# 5-5-5	# 1	# 5
10 min delay					
2 days delay					

Note. Yellow shades refer to the massed condition, blue to graduated intervals, red to uniform intervals, and purple and white to the control groups with one rehearsal only.

Test conditions:

a. 52 word pairs (i.e., benison – blessing) – 36 tested/ 16 fillers
b. Word pairs were divided into six conditions (36:6= 6 word pairs per condition)
c. Items were presented as: study-test-test-test (# denotes the first encounter as a study trial.)
d. The items were presented for 8 seconds each.
e. The response time was measured (slow= higher difficulty level= more effort= higher retention).
f. Participants were 48 students at Loyola University (four subjects were assigned to each counterbalancing condition).

A testing list with 121 items was compiled and all conditions occurred six times (see graph below for a possible test schedule).



Note. Every cell represents a test trial of 8 seconds duration. The pattern is continued to a total of 121 items. The numbers refer to the completed cycles of one condition (e.g. shades of yellow= massed condition)

Table A6: Interval patterns for Karpicke and Roediger's third test series
Karpicke & Roediger (2007b) – Experiment 3

Experiment 3 is described in detail because Karpicke and Roediger now introduce 4 encounters in order to create a delayed first testing condition. The table below describes their test scenario.

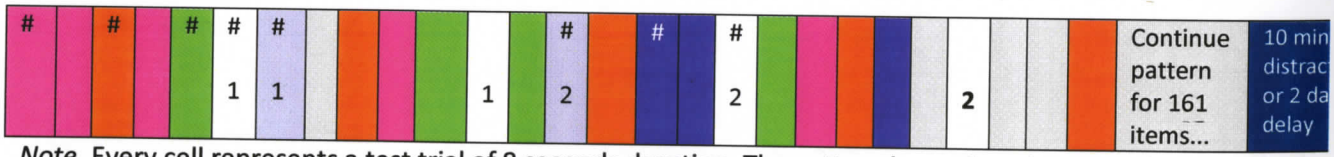
	# 0-1-5-9	# 5-1-5-9	# 5-5-5-5	#0-5-5-5	# 0	# 5
10 min delay						
2 days delay						

Note. Brown and dark purple shades refer to the massed condition, pink and green refer to graduated intervals, light purple and white to the control groups with one rehearsal only, and gray are filler items.

Test conditions:

a. 56 word pairs (i.e., benison – blessing) – 42 tested/ 14 fillers
b. Word pairs were divided into six conditions (42:6= 7 word pairs per condition)
c. Items were presented as: study-test-test-test (# denotes the first encounter as a study trial.)
d. The items were presented for 8 seconds each.
e. The response time was measured (slow= higher difficulty level= more effort= higher retention).
f. Delayed first testing was introduced as a variable.
g. Participants were 56 students at Loyola University (four subjects were assigned to each counterbalancing condition).

A testing list with 161 items was compiled and all conditions occurred six times (see graph below for an example of how the test items could have been arranged on this list).



Note. Every cell represents a test trial of 8 seconds duration. The pattern is continued to a total of 161 items. The numbers refer to the first–or second–completed cycle of a condition (#0 and #5).

Appendix B

Word frequency of target items in the textbook

The following list was compiled to document the incidental encounters of those target items in the textbook Deutsch NaKlar that were studied with ViVo and tested in the print vocabulary quizzes.
The numbers refer to the frequency in the chapter where this item was introduced as new vocabulary. Derivatives and compounds were included in the count if they were very close to the same semantic concept (i.e., das *Kleid* [dress] and Abend*kleid* [evening dress] were added, whereas *Kleid*ung [clothing] was not). Furthermore, verbs with stem vowel changes were included (i.e., nehmen, nimmst [to take]).

Table B1: Chapter 1 word frequency in the textbook Deutsch NaKLar

1	old	alt	9
2	to work (verb)	arbeiten	16
3	not	nicht	16
4	when	wann	3
5	here	hier	19
6	the name	der Name	24
7	the university	die Universität	3
8	to telephone	telefonieren	2
9	interesting	interessant	11
10	the music	die Musik	8
11	large, tall	groß	27
12	to be	sein (bin, bist, ...)	145
13	the friend (male)	der Freund	30
14	the book	das Buch	7
15	how	wie	53

16	often	oft	14
17	to come (verb)	kommen	50
18	good, well	gut	21
19	nice	schön	3
20	athletic	sportlich	3

Table B2: Chapter 2 work frequency in the textbook DNK

1	again	wieder	7
2	nothing	nichts	5
3	the television	der Fernseher	4
4	the kitchen	die Küche	14
5	the apartment	die Wohnung	41
6	the bed	das Bett	13
7	to eat	essen (isst)	24
8	to buy	kaufen	8
9	to run, jog	laufen (läuft)	18
10	to take	nehmen (nimmst)	17
11	to see	sehen (siehst)	21
12	comfortable	bequem	7
13	free	frei	6
14	dark	dunkel	3
15	the photo	das Foto	2
16	only	nur	8
17	beautiful	schön	14
18	expensive	teuer	7
19	the room	das Zimmer	101
20	to drink	trinken	5

Table B3: Chapter 3 word frequency in the textbook DNK

1	through	durch	9
2	him	ihn	11
3	me	mich	4
4	without	ohne	8
5	my	mein	110

6	the parents	die Eltern	27
7	the present	das Geschenk	13
8	to give	geben (gibst)	17
9	to congratulate	gratulieren	4
10	to become	werden (wirst)	25
11	third	dritte	2
12	new	neu	16
13	first	erste	9
14	the son	der Sohn	10
15	to celebrate	feiern	26
16	to wish	wünschen	5
17	seventh	siebte	1
18	important	wichtig	9
19	for	für	19
20	against	gegen	5

Table B4: Chapter 4 word frequency in the textbook DNK

1	the coffee	der Kaffee	2
2	to call/phone	telefonieren	8
3	to meet	treffen (trifft)	6
4	to must, to have to	müssen (muss)	26
5	late	spät	11
6	Thursdays (adverb of time)	donnerstags	4
7	the afternoon	der Nachmittag	13
8	the week	die Woche	25
9	the plan	der Plan	13
10	the minute	die Minute	5
11	to be allowed	dürfen	12
12	to shop	einkaufen	15

13	to be able to	können (kann)	23
14	early	früh	7
15	to take along	mitnehmen (nimmt ... mit)	4
16	to go along, join	mitkommen	8
17	to have breakfast	frühstücken	3
18	the fitness center	das Fitnesscenter	1
19	to want	wollen (will)	23
20	the hour	die Stunde	13

Table B5: Chapter 5 word frequency in the textbook DNK

1	me	mir	23
2	yellow	gelb	5
3	with	mit	28
4	fashionable	modern	1
5	to, after (town)	nach	15
6	from, out of	aus	23
7	fresh	frisch	7
8	the bread	das Brot	11
9	(the) vegetable	das Gemüse	10
10	the dress	das Kleid	8
11	the sugar	der Zucker	2
12	the salt	das Salz	1
13	the beer	das Bier	4
14	to fit	passen	13
15	to be pleasing	gefallen (gefällt)	17
16	to wear	tragen (trägt)	27
17	to show	zeigen	13
18	to pay	bezahlen	2
19	to help	helfen (hilft)	8
20	the potato	die Kartoffel	4

Appendix C

Test corpora practiced with ViVo and their German word frequency ranking

The frequency ranking for the test corpora is based on the German frequency dictionary (Jones & Tschirner, 2006). It is based on 4.2 million words of Contemporary German. These had been compiled out of different corpora (spoken; literature; newspaper; academic texts; instructional language). The dictionary displays 4,034 of the most frequent words. The frequency calculation is based on how often this word appears per million. The value given in brackets indicates the authors' ranking. The lower this value is, the more frequent this word is in German.

If entries did not produce an exact match the next closest word was chosen if possible (i.e., zweite – not in; zwei = 77). The word researched in multi-word phrases is in bold)

Table C1: Chapter 1 Vocabulary

aber natürlich (32)	**gut (78)**	**machen (49)**	**sportlich (1902)**
alt (116)	**hier (71)**	**Musik (469)**	**stressig (Stress= 2566)**
arbeiten (200)	**hören (1557)**	**Name (270)**	**studieren (436)**
auch (16)	**immer (68)**	**nett (1438)**	**telefonieren (3352)**
Beruf (477)	**interessant (531)**	**nicht (12)**	**Universität (319)**
Buch (295)	**Jahr (51)**	**oft (215)**	**wann (583)**
danke (778)	**Karten spielen (1191)**	**ruhig (905)**	**wer (173)**
Freund (327)	**kommen (61)**	**sein (3)**	**wie (28)**
Geburtstag (1766)	**lernen (203)**	**Semester (1342)**	**Wohnort (wohnen=380)**
gehen (69)	**lesen (323)**	**Spaß (666)**	
groß (74)			

Table C2: Chapter 2 Vocabulary

Bad (1577)	Fernseher (2415)	laufen (248)	teuer (936)
bequem (2456)	Foto (1396)	nehmen (139)	trinken (608)
Bett (654)	frei (318)	nichts (111)	Uhr (349)
billig (1672)	Geld (249)	nur (44)	viel (60)
da (35)	hell (1397)	Regal (2640)	warum (246)
dunkel (819)	kaufen (581)	schön (164)	weit (122)
Durst haben (not in)	kein (50)	schreiben (245)	wieder (75)
essen (655)	klein (114)	sehen (81)	Wohnung (418)
etwas (107)	Küche (960)	so (21)	Zeit (90)
fahren (169)	Lampe (3894)	suchen (293)	Zimmer (609)

Table C3: Chapter 3 Vocabulary

Bruder (687)	geben (57)	mein (53)	Tochter (514)
dich (212)	gegen (117)	mich (67)	Tradition (1344)
dritte (Drittel= 1844)	Geschenk (2610)	Mutter (227)	um (47)
durch (56)	gratulieren (not in)	neu (80)	Vater (216)
Eltern (351)	Großvater (3457)	ohne (119)	Weihnachten (2726)
erste (91)	heiraten (1298)	Party (3534)	werden (9)
feiern (1059)	ihn (93)	planen (499)	wichtig (177)
Frau (103)	Kalender (not in)	Schwester (776)	wünschen (684)
fünfte (fünf= 272)	kennen (181)	siebte (sieben= 570)	zehnte (zehn= 214)
für (18)	leider (642)	Sohn (448)	zweite (zwei= 77)

Table C4: Chapter 4 Vocabulary

Abend (313)	früh (322)	mitkommen (not in)	samstags (1306)
anrufen (1386)	frühstücken (Frühstück= 2681)	mitnehmen (1729)	spannend (2118)
Bibliothek (927)	gemütlich (3451)	mittags (2194)	spät (171)
doch (72)	halb (411)	morgens (311)	Stunde (262)
donnerstags (1332)	*ins* Kino *gehen* (1825)	müssen (45)	Tasse (3503)
dürfen (142)	Kaffee (1420)	Nachmittag (1426)	treffen (287)
einkaufen (1789)	können (23)	nachts (335)	vor (55)
fernsehen (1060)	Konzert (1721)	Oper (2552)	vormittags (3685)
Film (524)	lieber (754)	pro *Woche* (589)	Wochenende (764)
Fitnesscenter (not in)	Minute (361)	Plan (987)	wollen (65)

Table C5: Chapter 5 Vocabulary

aus (41) Apfel (3837) bei (29) Bier (1713) blau (1016) braun (2178) Brot (1757) danken (2667) Fleisch (1624) frisch (1297)	gefallen (536) gelb (1819) Gemüse (3450) glauben (143) grau (1526) helfen (406) Hose (2422) Kartoffel (3609) Kleid (1681) mir (64)	mit (13) modisch (modern= 504) nach (38) orange (not in) passen (775) Rock (3225) rot (381) Salz (2437) schmecken (2373) Schuh (2114)	seit (141) Tasche (1638) tragen (305) von (11) wem (not in) wohin (1837) zahlen (965) zeigen (154) zu (6) Zucker (3823)

Appendix D

German Placement Questionnaire A

Name:

Faculty:

Majoring in:

E-Mail:

Year:

Did you take German in High School?

Where?

For how many years?

Did you take a German course at a university/college before?

How much time have you spent in a German-speaking country?

When was the last time you were in a German-speaking country?

Do you have German-speaking heritage? If so, in what way?

Have you ever learned any other languages? Which?

Appendix E

Student Questionnaire B

Name of German 100A Instructor: _____
(Please note: Your instructor will not have access to this questionnaire)

Your name: _____ date: _____

1. What did you find most helpful to remember the words:
   - 0 the images
   - 0 the sound
   - 0 the sample sentences
   - 0 the fact that you could type in the words for practice.
   - 0 other, please describe: .....

2. In general, do you like to learn by:
   - 0 visualizing things
   - 0 listening to things
   - 0 practicing/ doing things

3. Do you prefer learning words with:
   - 0 a textbook
   - 0 an online program
   - 0 both
   - 0 other, please describe: .....

4. Do you prefer to study:
   - 0 alone
   - 0 in pairs
   - 0 in groups

5. Please check all that apply: Do you feel comfortable using:
   - 0 computers
   - 0 Moodle
   - 0 ViVo
   - 0 none of the above

6. How much do technical issues with computers or online applications bother you:
   - 0 not at all
   - 0 a little
   - 0 much
   - 0 very much

7. How often did you study the vocabulary on your ViVo word list:
   - 0 every day
   - 0 on ViVo days
   - 0 just before the practice quiz
   - 0 just before the marked quiz

8. How often did you study the other vocabulary in the textbook:
   - O every day
   - O just before the practice quiz
   - O just before the marked quiz
   - O I did not work out a schedule
   - O never
   - O other (Please describe) .....
     .....

9. At what time of the day do you prefer to study?
   - 0 morning
   - 0 afternoon
   - 0 evening
   - 0 night
   - 0 It doesn't matter, because .....
     .....

10. On a scale from very easy to very difficult how would you rate your ability to follow the practice schedule of ViVo:

very easy .....very difficult

0 1 2 3 4 5 6 7 8 9 10

What made it easy/difficult? Please explain:

11. From one ViVo practice to the next - how well did you remember the words? Please rate your effort:

It was easy I remembered all relearn most — It was difficult I had to

0 1 2 3 4 5 6 7 8 9 10

12. If you skipped a ViVo practice – you did so because:

0 you felt you did not need another one
0 it did not fit into your schedule
0 you forgot

13. Did you think some encounters on ViVo were more important than others?
0 No, I needed all 5 encounters
0 I could skip encounter number _____ because.....
.....
.....

14. What was your favourite part about ViVo?

15. What would you do – if anything – to improve ViVo?

16. Other strategies to learn vocabulary. Choose one favourite and rate the others
17.

		I often do this	I never do this	I seldom do this	I always do this	my favourite
1	I use a dictionary					
2	I colour code grammatical features					
3	I test myself with word tests.					
4	I write flash cards and carry them with me to study					
5	I write word lists and repeat the words many times					
6	I say the word out loud when I am studying					
7	I keep a notebook with the words I want to learn					
8	I label objects with stickers					
9	I need to write the word a couple of times					

Appendix F

Error protocol

The table below represents an error protocol of 6 online quiz vocabulary items. These items were chosen based on their frequency in the textbook.

The answers of all students for these 6 words were tracked in the Moodle online quiz user log. The percentage values indicate how often this word was misspelled or not produced. Furthermore, the type of error is documented.

Table F1: Incidental encounters —Moodle error protocol

frequency	8	7	2		53	28	41
UG	kaufen = no answer	7% Buch 7% der Buch 3% buch 3% die Buch	9% die Zucker		wie = no answer	4% von	11% der Wohnung 11% das Wohnung 6% der wohnung 6% das Zimmer
GG	4% kufen	7% der Buch 3% das Buchen 3% die buchen 3% buchen 3% die Buch	16% die Zucker 10% das Zucker		3% wihr	3% bei	15% der Wohnung 8% das Wohnung 4% die Appartement 4% das Appartement 4% die Wohnug 4% das Apartment
correct response	kaufen	das Buch	der Zucker		wie	mit	die Wohnung