Language Frequency Profiling of Written Texts by Students of German as a Foreign Language

by

Bjanka Pokorny

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Arts
in
German

Waterloo, Ontario, Canada, 2009

AUTHOR'S DECLARATION

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

The present work contributes to the ongoing discussion of the factors involved in perfecting foreign language learning through a close examination of vocabulary use. Motivated by Laufer's (1991) argument that the use of less frequent vocabulary items is a sign that a language learner is approximating the lexical competence of a native speaker, I set out to model Laufer and Nation's (1995) study that assessed lexical frequency. The first goal of this work was to assess the usefulness of the lexical frequency profile (Laufer and Nation, 1995) in evaluating written texts produced by learners of German. This lexical frequency profile had mostly been used to examine vocabulary use of learners of English. Instead of using frequency bands of German, this work relied on three generated word frequency lists. The second goal of this work was to examine how the language repertoire of aspiring bilinguals varies at the lexical level by comparing vocabulary use at three competency levels (Introductory German I, II and Intermediate German). The analysis revealed that the lexical frequency profile is a valuable tool for evaluating lexical use by language learners, although the tool was difficult to adapt for research of texts in German. Furthermore, learners in all three courses relied heavily on vocabulary from learning materials used in their courses, and they were more likely to use less frequent words as they progressed from the introductory to the intermediate language course.

Acknowledgements

First and foremost, I would like to thank my supervisor Mathias Schulze. Without his encouragement, guidance and endless source of ideas and suggestions, this project would have never been completed.

I would also like to express my gratitude to Peter Wood and Grit Liebscher for their willingness to share their data and materials with me so readily. Thank you both very much. I would furthermore like to thank Kaitlyn Kraatz for suggestions on earlier drafts of this work.

The endless amount of support I received from my parents while working on this project cannot be overlooked and for that support I am eternally grateful. I must also thank my sister for visiting me regularly with the kids and thereby forcing me to take a break from my thesis. Teodora deserves a big thank you for just being Teodora.

I must also acknowledge the help I received from others around me: Vlado in particular, possibly the biggest clown of all times, must be thanked for regularly making me laugh during the past few months. A big thank you goes out to all of my ladies – far and near: Jokićke, Kliewer, Vio, my Mannheimer Blonde Squad, M.S.K., Olga, & Ellen. Thank you.

Additionally, I would like to thank Ali & Laura: As stressful as they may have been, I will always cherish these past few months we have spend together in the offices, working on our theses as diligently as rehearsing our now (in)famous dance routines. The modern language building and Ring Road have seen some great times because of us. Benny Lava.

Lastly, I would like to my committee members Barbara Schmenk and Sarah Turner for their suggestions for and improvements to this work.

Dedication

For my family.

Table of Contents

List of Figures

Chapter 1
Introduction

> "Without Grammar, very little can be
> achieved. Without vocabulary,
> nothing can be achieved." (Wilkins,
> 1972, III)

Whether it is done for personal interest or professional gains, learning a foreign language is difficult work. Balancing semantics while simultaneously juggling lexis can be overwhelming. Although many succeed in walking this tightrope of language learning, researchers can still not claim to fully understand the entire process. In fact, many would agree that only a small portion of this artistic act can be explained thoroughly.

The present work contributes to the ongoing discussion of the factors involved in perfecting foreign language learning. This was accomplished by joining a growing group of researchers (see: Laufer and Nation, 1999; Davidson, Inderfey and Gullberg, 2008; Kim, 2008; Zareva, Schwanenflugel and Nikolova, 2005; Daller, Van Hout and Treffers-Daller, 2003) that focus on vocabulary acquisition in second language learning and are united in the belief that vocabulary acquisition is a central component of language learning (Read and Chapelle, 2001; Huibregtse, Admiraal and Meara, 2002, Hover, 2003). After all, words represent the basic ideas speakers want to express and vocabulary knowledge has been identified as a potential predictor of the overall proficiency a language learner will achieve

(Zareva, Schwanenflugel and Nikolova, 2005; Kim, 2008). In fact, Zareva, Schwanenflugel and Nikolova (2005) argue that a strong vocabulary foundation in the target language serves as the strongest predictor of the proficiency a learner can achieve in this language. Basanta (2004) and Crossley (2009) add that speaking and reading skills are also greatly enhanced by improvements in one's vocabulary and that lexical growth correlates strongly with academic achievement.

These arguments strongly guide the approach and goals of this work as they underline the urgency of understanding vocabulary development and acquisition in as much depth as possible. Vocabulary acquisition and use is a particularly opportune research field and it has been considered in combination with many other linguistic phenomena such as grammar or fluency (Larsen-Freeman, 2006). This process is also frequently studied in isolation, such as in studies that attempt to model or quantify a learner's vocabulary knowledge (Meara, 2004). Although the knowledge generated has deepened our understanding of both vocabulary acquisition and development across time, the research field continues to allow for constructive and motivating research questions. The present work will focus on only one subdivision of the research field, namely vocabulary use in learners' written texts.

1.1. Previous Research on Vocabulary

Before addressing the present research question in more detail, previous reflections on vocabulary research must be reviewed. The question of what constitutes having knowledge of a word has often been debated and several conceptualizations have been offered. Nation

(2001) is in favor of a detailed approach and cites four criteria that, although difficult to measure, encompass lexical knowledge: form is defined as the learner's knowledge of the spoken and written form of the word. Position encompasses knowledge about the grammatical behavior of the word. Function evaluates whether the learner is familiar with the word's frequency and appropriateness within registers. Lastly, meaning addresses the learner's knowledge about the conceptual content of the item and his word associations.

Leach and Samuel (2007) cite form, meaning and syntactic roles as the dimensions associated with knowing a word. Form includes the phonetic and orthographic information associated with the word. Meaning for these authors is a multidimensional and context-specific variable. Syntactic roles refer to additional information associated with the item such as factual information about the associated concept that allows the speaker to use the item appropriately. Others opt for a more straightforward definition. Corrigan (2007) argues that vocabulary knowledge is the ability to use a word in a novel context and construct meaning from text. The examples given above present only a few frameworks offered as an explanation for knowing a word, but illustrate well how diverse these frameworks can be. A more detailed discussion is available in Dewaele (2008), Wolter (2001), Crais (1990), and Elman (2004).

Like Corrigan, Read and Chapelle (2001) argue that vocabulary knowledge should be equated with vocabulary use and is best examined in context. This thesis project relies on the approach offered by Corrigan (2007) as the only units of analysis taken into consideration are students' written submission without consideration for their metacognitive knowledge about the words used.

While it is theoretically possible to define what it means to know a word, designing a measure that captures the degree of vocabulary knowledge in all its facets is significantly more challenging (Zareva, Schwanenflugel and Nikolova, 2005). The authors also argue that it is unclear what differences exist between the knowledge of a word by a native speaker and knowledge of the same word by a non-native speaker. As such, measures established for one population are difficult to use for and apply to another. This argument further shapes the course this work will pursue as the units of analysis are words used in written texts by learners of German. The focus here will not be to assess how well a certain set of words is known by a learner but rather the focus will be to assess the state of learners' vocabulary as they use it in written texts. In other words, this thesis will not focus on learners' metacognitive abilities or their vocabulary acquisition processes but rather exclusively on the way they use vocabulary and the type of vocabulary they present in their written texts.

The majority of research carried out in L2 vocabulary studies has focused on developing measures to assess proficiency or the size of a learner's vocabulary (Davidson, Inderfey and Gullberg, 2008). Doing so has been accomplished through various measures ranging from interview methods to paper and pencil measures (Bachman, 2000). The latter are regarded as more reliable as they avoid many validity concerns associated with face-to-face testing. These concerns range from failure to address all vocabulary known to the student, to the impossibility of differentiating between a learner's capacity to use a word and his ability to simply recognize it upon hearing it (Zareva, Schwanenflugel and Nikolova, 2005; Kim, 2008).

Pencil and paper measures of vocabulary are also not without fault. Laufer and Nation (1995) provide an overview of such measures for vocabulary size along with the research problems with which these measures are associated. To summarize, these range from great sensitivity to text content or length, to equally severe validity concerns. Webb (2008) adds that the majority of vocabulary tests are heavily biased towards the measurement of receptive vocabulary size, which generally exceeds the size of a learner's productive vocabulary.

While receptive vocabulary knowledge is important for language learning as it has been argued that receptive knowledge of the 2000 most frequent word families allows for the understanding of 90% of the words in spoken discourse (Webb, 2008), the focus of this work is on a different dimension of vocabulary. The objective of this work is not to assign numeric values to learners' vocabulary sizes. Rather, the goal of this work is to examine the quantitative properties of the vocabulary used. The question steering this project is whether or not students rely solely on the vocabulary present in their textbook or also use vocabulary that was not printed in the learning materials. Evaluating how much students rely solely on vocabulary taught through their textbook is of interest to researchers and educators who invest time and effort in teaching students a foreign language and also to those creating learning materials and learning aids.

Before proceeding with a more detailed description of the present work, I will discuss past research on vocabulary use. A literature review revealed variations in how native and non-native speakers learn and use the vocabulary of a given language. These findings will be discussed in the following section.

Chapter 2

Language Knowledge and Use in Native and Non-Native Speakers

The way native speakers use their mother tongue varies greatly. Not all author moving literature or timeless poetry, but all succeed in communicating with others. Similar variability is also readily observable among non-native speakers. Few non-native speakers acquire the foreign language well enough to be mistaken for native speakers; yet most non-native speakers possessing some proficiency succeed in communicating with others.

Related frameworks can be used to explain the variability of linguistic talent in both native and non-native speakers. Auer and Bernstein (2008) write that an individual adult's lexical knowledge is the result of that person's psycholinguistic experiences in interaction with biologically determined language processing factors. An addition to this argument is needed to account for the linguistic success of foreign language learners. For non-native speakers, effort invested in and study towards further vocabulary acquisition can be added as one component to explain the success language learners attain.

Differences and similarities between native and non-native speakers exist when we examine how a lexical item is learned. Adding a new word to one's vocabulary consists of more than simply learning the sound of the item. Acquisition of a lexical item entails embedding the item in a lexical network – connections between semantically related items. This creates an idea of vocabulary learning as system learning rather than individual item learning (Schoonen and Verhallen, 2003). Given that this knowledge evolves over a long period of time, it can be thought of as static at any given moment. Put differently, if one's vocabulary knowledge were assessed on a weekly basis, little difference would be expected during such short intervals.

The process of learning a lexical item in a foreign language builds on the above outlined process in the following way: during early stages of lexical learning, an item generally contains only phonological and/or orthographic information. This is understandable as language learners are normally already familiar with the concept the entry represents. Exceptions to this include novel, culture specific concepts that learners might encounter such as customs or traditions (Dewaele, 2008).

Additional stages of acquisition include the transfer of semantic information from the native language. Therefore, the resulting semantic associations for the item are a combination of the foreign language form and the native language semantic information (Corrigan, 2007). The semantic networks available in one's native language support and enhance the acquisition of a foreign language by aiding the learner in categorizing newly acquired words, indicating that the lexical knowledge possessed in the native language has a strong influence not only on vocabulary acquisition but also on target language lexical formation (Ijaz, 1986).

Moreover, the lexical conceptual knowledge of one's native tongue has a strong influence on lexical knowledge in the target language (Brent, 2006). While the native and target language rarely overlap perfectly in terms of grammatical structure or word order, Brent (2006) writes that similarities between one's native tongue and the target language often raise the likelihood that successful acquisition will occur.

Although research within this domain is increasing in size and breadth, few findings have been established as facts. For this reason, Fitzpatrick (2007) cautions against establishing native speaker norms for vocabulary knowledge and use. The author further writes that most studies attempting to establish such norms rely on vocabulary tests created with high frequency words in a given language. Once less frequent words are used in the tests, homogeneity disappears and responses of native speakers show great variation.

A similar caution has been made regarding non-native speakers norms. While *norms* will not be discussed here, previous research does allow for a discussion of broad findings regarding lexical knowledge.

One noteworthy difference exists between the acquisition of lexical items for native and non-native speakers. Native speakers acquire their early words through spoken language. As their literacy increases a larger percentage of words are acquired through print (Auer and Bernstein, 2008). This pattern differs from vocabulary acquisition in a foreign language classroom that normally relies heavily on printed instruction materials right from the start.

Relying on data from word recognition tasks for both native speakers and non-native speakers, Auer and Bernstein (2008) write that words acquired earlier are associated with faster and more accurate performance than words acquired later. Put differently, items that have been known longer have been integrated better into semantic networks and are for that reason more easily retrievable. Corrigan (2007) adds that knowledge of lexical organization is crucial for deep vocabulary knowledge – defined as the presence of complex connections to central concepts – because words with more complicated sets of connections to other words will tend to be known more deeply than those with more tenuous connections. This holds true for native and non-native speakers.

Another similarity native and non-native speakers share relates to the frequency of words in one's vocabulary. Researchers agree that the more low frequency words are known by a speaker the larger the known vocabulary is (Laufer and Nation, 1995; Laufer, Elder, Hill and Congdon, 2004; Alderson, 2007). This argument pertaining to lexical frequency further steers the aim of this work.

Given these differences and similarities in lexical frameworks between native and non-native speakers, conclusions can be drawn about the influence these characteristics exert on written texts produced by these two groups.

Learners writing in the target language devote a lot of attention to the vocabulary they use (Ellis and Yuan, 2004). This concentration on translation often occurs at the expense of fluency, thus resulting in shorter texts when the two groups perform under time constraints.

Because learners have a smaller vocabulary in the foreign language in comparison to their native tongue, they use more words of general rather than specific meaning. This dependency on general vocabulary items leads to sometimes inappropriate overgeneralizations, incorrect collocations and disregard for register (Crossley, 2009).

Schoonen and Verhallen (2003) write that texts produced by native speakers and non-native speakers differ remarkably not only in fluency but also in bursts of text production. Linguistic variables play a bigger role in foreign language writing, forcing students to devote more attention to vocabulary. In native language writing, individuals rely more on their metacognitive skills which allow one to spend less time focusing on vocabulary and grammatical aspects and attend more to fluency. For non-native speakers, this effect is mediated by experience with the target language as more exposure to the language is associated with increased bursts of text production.

Texts produced by non-native writers are also characterized by a less complex structure (Schoonen and Verhallen, 2003; Ellis and Yuan, 2004; Tavokoli and Foster, 2008). Learners of a foreign language encounter lower order problems of word finding and grammatical structure. As this requires too much conscious attention it leaves little or no working memory capacity free to attend to higher level or strategic aspects of writing such as organizing the text in a more complex fashion or trying to convince the reader of a certain view. This trade-off between accuracy and complexity is reduced as proficiency increases (Ellis and Yuan, 2004).

The above findings further serve to illustrate the great significance of vocabulary in the study of foreign language learning and how the extent of vocabulary knowledge can influence text production. The following section will present a more detailed discussion of a specific type of vocabulary, namely high frequency vocabulary that is of great relevance to this work. The implications and use of high frequency vocabulary items and their importance for this work will be discussed below.

2.1. On the Importance of High Frequency Vocabulary Items

While vocabulary itself is an essential aspect of language learning, be it foreign or one’s own, researchers agree that not all words are equally important or easy to learn. One defining feature is the frequency of the lexical item (Alderson, 2007). We all encounter some words more often than others. We also all use some words more than others, and some words in our native tongues we never use at all.

Words we hear, use and encounter regularly are often high frequency words. Alderson (2007) writes that native speakers are capable of correctly estimating word frequencies in their native languages. He adds that their perceptions of word frequencies adapt to their increasing vocabularies over time. Put differently, as we add more words to our lexicons, we also become more aware of the frequencies of these words. We seem to be well equipped to see patterns in word frequencies in our native tongues.

To go one step further, many of us can say words like "*hello*", "*yes*" or "*no*" in a language we have never formally studied. Again, these are frequent lexical items and would have been encountered with higher frequency than items like "*exhaustion*", "*sleeplessness*" or "*master's thesis*". The rationale for this argument is that words in a foreign language are generally learned in order of their frequency. For this reason, words like "*yes*" and "*no*" might have been encountered in casual conversation with a native speaker or through exposure to a medium in that language.

When the foreign language is studied in a formal setting such as a language classroom, the same rule for frequency applies. Textbooks and courses are designed to foster more frequent vocabulary in favor of rare vocabulary because the time spent teaching rare words is believed to outweigh the benefits associated with knowing such words.

Moreover, high frequency items, in foreign and own languages, need to be known for minimal comprehension of most written texts (Alderson, 2007) as items of high frequency frame most of our daily interactions.

But reliance on and sole use of high frequency items are the result of a more limited vocabulary size (Basanta, 2004). Items of low frequency – within both the native and foreign language– are indicators of a larger vocabulary. Thus the larger one's vocabulary, the more low frequency items are known. Going one step further, Laufer (1991) argues that the acquisition of low frequency items in a foreign language is a sign that the learner's vocabulary is approximating the lexical competence of a well-spoken native speaker. This finding plays a large role in the course this thesis project will take.

Many arguments and findings have been discussed so far in this work. I discussed how vocabulary had been measured in the past as well as how vocabulary use differs greatly based on a learner's proficiency. Although they are already enlightening on their own, it is the combination of the aforementioned findings that has captured my interest. Through this project, I set out to examine two issues. Dissatisfied with other measures of vocabulary, I sought to find one that would evaluate vocabulary use not by quantifying size but rather by describing the type of vocabulary learners use. I found one such measure in Laufer and Nation's (1995) work in which the authors introduce and put to use the lexical frequency profile – a framework that allows for the evaluation of vocabulary of different frequency bands found in learners' writings. Laufer and Nation (1995) have shown that the lexical frequency profile is useful and valid tool for the assessment of vocabularies of English learners. My first goal in this work is to appraise the usefulness of the lexical frequency profile (Laufer and Nation, 1995) – a design discussed in the following section – in evaluating written texts produced by students of German from three German language courses (Introductory German I, II and Intermediate German).

Secondly, inspired by Laufer’s (1991) argument that the use of less frequent vocabulary items is a sign that a learner is approximating the lexical competence of a native speaker, I wanted to examine how the language repertoire of aspiring bilinguals varies at the lexical level by comparing vocabulary use at three proficiency levels (Introductory German I and II and Intermediate German). In other words, this work will evaluate to what extent learners rely solely on the vocabulary they were taught in their language classes, or if they venture out and use vocabulary not acquired through the course materials.

Chapter 3

The Lexical Frequency Profile

In light of the measurement concerns associated with vocabulary assessment tools available, Laufer and Nation's (1995) lexical frequency profile aligns well with the purpose of this present work. The lexical frequency profile is a way of analyzing a text through two pieces of software, *Frequency* and *Range,* both of which will be discussed in more detail in subsequent sections. To summarize briefly here, the combination of the two pieces of software allow for the categorization of words into frequency bands against which a learner's writing is evaluated. The scores and the wordlists obtained through the analysis of a text provide insight to lexical variation in the analyzed text.

Laufer and Nation (1995) analyzed students' written texts by comparing them to three frequency lists: one containing the 1000 most frequent word families in the English language, one containing the 2000 most frequent word families in the English language, and a third list containing 500 academic words used at university. Word family within the authors' study is represented by a "head word" and all "derived forms" as in the example: push: pushed, pushes, pushing. The authors argued that their results indicate that that the use of frequently occurring words reflects a smaller vocabulary while the use of low-frequency words is an indicator of vocabulary richness. This measure does not identify whether a learner can produce a certain word when prompted to do so, but rather how much lexical variety he uses in his writing (Laufer, 2005).

Thus the lexical frequency profile does not quantify vocabulary size, but rather provides the percentage of words a learner uses at different frequency levels in writing. As more frequent words are crucial for effective communication and are therefore acquired sooner, this framework is suitable as it allows for a reliable estimate of how diverse the learner's written vocabulary is once more attention is paid to the less frequent items.

Laufer and Nation (1995) present the lexical frequency profile as an objective analysis of a learner's vocabulary. The authors applied this measure to learners of English and concluded that the lexical frequency profile is a reliable and valid tool to access students' lexical use in writing. Their analysis provided stable results for two pieces of writing by the same learner, discriminated between two pieces of writing by learners of different proficiency levels, and the results correlated with an independent measure of vocabulary.

The authors further argue in favor of the lexical frequency profile by saying that it is free of subjective decisions that could be encountered if the texts were analyzed by one or more instructors, allows for the discrimination between learners who use frequent and non-frequent vocabulary and between those who can and cannot vary their restricted vocabulary. Calculations of word frequency are carried out by two software programs - Range and Frequency - which further reduces the bias of this tool (Heatley, Nation and Coxhead, 2002). The same software will be used in this project and will be discussed at a later point. Furthermore, Meara (2005) writes that there is general agreement as to which words should appear in which frequency list, making this tool less dependent on the context or genre of the written text.

The lexical frequency profile was administered for research purposes primarily to students learning English as a foreign language. To the best of my knowledge, only two studies focused on students of German (East, 2004; East, 2006). Both of these studies investigated specific aspects of course design such as the benefits associated with dictionary use during tests, which are not of interest to this project. Rather, the goal of this work is to use the lexical frequency profiling approach in order to learn more about the student population of three different German language courses offered at the University of Waterloo. In doing so, I will be able to identify trends in the learners' lexical use which will yield a better understanding of their learning processes.

Laufer and Nation's (1995) application of this framework relied on word frequency bands for the English language. Because identical lists for the German language could not be found, new lists were generated for the present work. The lists used for comparison in this work were the *Vorsprung* Vocabulary List (henceforth *Vorsprung* list), Wiktionary's list of the 2000 most frequently occurring words in subtitles (Subtitles vocabulary list) and a final list containing the 10, 000 most frequent words of the German language as created by *Projekt Deutscher Wortschatz* at the University in Leipzig (*PDW* vocabulary list).

Given that the three lists were either generated or modified from their original state to fit the constraints of this study, they cannot be taken to represent frequency bands for the German language like the lists that Laufer and Nation used for the English language. If considered outside of this work, these lists are better described as containing different genres of language use. However, each of the lists contains the most frequent words for the genre it represents.

For the purpose of this work, these three lists will be used like frequency bands. Additionally, the items within each list are not arranged in order of frequency but rather alphabetically. Each of these three lists will be discussed below.

3.1. *Vorsprung* Vocabulary List

This list was generated though the compilation of all lexical items from the textbook used for the Introductory German I, II and Intermediate German (German 101, 102 and 201 in that order) courses. The textbook for all three courses was *Vorsprung: An Introduction to the German Language and Culture for Communication* (Lovik, Guy and Chavez, 2007), and this list was compiled by including all words students were taught while using the book.

The inclusion of vocabulary lists containing basic and frequent vocabulary, such as the *Vorsprung* list used in the present study, is criticized by Daller, Van Hour and Treffers-Daller, (2003). The authors caution against the use of vocabulary lists containing such basic and frequent vocabulary. They argue that such items do not provide size estimates because of their strong and repetitive nature, but rather only serve to contaminate the results.

The authors' argument aligns with one of the questions this work aims to answer, namely whether students' vocabulary use changes as they progress in language courses. The *Vorsprung* list was included in the analysis as it covers vocabulary students learned formally in the classroom and is therefore familiar to all students in the three datasets. The inclusion of this list also allows for the calculation of what proportion of the students' written texts consist of vocabulary taught in their language courses.

Again, no claim can be made that this list features the most frequent words of the German language in all areas of use. On the contrary, some of the topics and vocabulary that *Vorsprung* covers such as *Grammatik und Strukturen* [grammatical structures] and *Märchen* [fairy tales] would rank relatively low if considered with reference to everyday use of German. However, given that this list contains vocabulary presented in the textbook students use for their course, one can deduce the words on this list have been encountered more frequently by the students than words in the additional two lists. In other words, this list contains the most frequent words the students encountered by being enrolled in the courses at hand in this work. A more detailed discussion on the descriptive characteristics of this vocabulary list and its similarity to the other two vocabulary lists will be presented in section 4.4.1.

3.2. Wiktionary’s List of the 2000 Most Frequently Occurring Words in Subtitles

This list was generated from subtitles of movies and television series with a total of about 25 million words in 2009 (Wiktionary, 2009). The use of this list in the present study was deemed appropriate as it reflects more colloquial language use that learners might have encountered if they had sought out additional mediums of the German language other than the materials used in the course.

This list covers a variety of semantic fields that would not be presented in a language textbook. In contrast to the *Vorsprung* list and the list generated by *Projekt Deutscher Wortschatz*, this list contains slang words and occasional English items frequently used in the German language.

3.3. *Projekt Deutscher Wortschatz*

The final list contains the 10, 000 most frequent words of the German language and was generated by *Projekt Deutscher Wortschatz*. Developed in 1995, *Projekt Deutscher Wortschatz* is a linguistic database created and funded by the Natural Language Processing Group (Automatische Sprachverarbeitung) within the Department of Computer Science at the University of Leipzig, Germany (Quasthoff and Wolff, 1999).

Electronic newspapers including but not limited to *Abendblatt, Berliner Zeitung, Die Zeit, Spiegel Online, Telepolis, Westfalenpost, Welt, Neues Deutschland* and *ZDF Heute* are the primary source of data for this database (V. Boehlke, personal communication, November 24, 2008). Additional data for this database and its word list are accumulated through a variety of electronic sources such as subject specific journals and newspapers on topics including but not restricted to medicine, law and computer studies (Quasthoff and Wolff, 1999).

Since 1995 *Projekt Deutscher Wortschatz* has accumulated a German text corpus of more than 500 million words with approximately nine million different word forms in an estimated 36 million sentences (Biemann, Bordag, Heyer , Quasthoff and Wolff, 2004). Aside from offering large volumes of data, *Projekt Deutscher Wortschatz* also allows for the extraction of information on basic syntactic and semantic information about a word as well as sentence-based word collocations and frequency occurrences of individual words (Biemann, Bordag, Heyer, Quasthoff and Wolff, 2004; Faulstich, Quasthoff, Schmidt and Wolff, 2002).

The data offered by *Projekt Deutscher Wortschatz* were selected for this project for several reasons. The primary appealing feature is the wealth of data and the ease of access to these data. The sample used in this work was based on a list entitled “The 10,000 most frequent words in the German language”. The list was modified slightly for this work and these modifications will be discussed in a subsequent section.

As mentioned above, a variety of widely read print sources was used to compile this list. In other words, the items in this vocabulary list cover a large selection of topics and semantic fields. The utilization of this list therefore serves two purposes. Firstly, once compared with the *Vorsprung* vocabulary list (see section 4.4.1). Preparation of all materials used in the study; a large number of words were found to overlap between these two lists. This overlap serves as a validating feature for the *Vorsprung* vocabulary list as it illustrates the high number of frequent words in the *Vorsprung* vocabulary list. Secondly, this list is the largest one used in this study. Given the size and source of this list, I argue that this contains a high number of low frequency words as well, something that cannot be said for the previously introduced two lists. Thus the utilization of this list allows for the examination of how many low frequency vocabulary words students have. This could be used to examine if students would be able to read some of the print sources that were used when this list was compiled.

In addition, Davidson, Inderfey and Gullberg, (2008) write that the more frequent words from linguistic corpora are also more likely to occur in textbooks. However, for the purpose of the present work it was deemed inappropriate to rely solely on the words that are used in the textbook as it could not account for any items used that do not occur in the book.

Having introduced the objective of this work and the materials that will be used to complete the analysis, a discussion on the course design will be presented to illustrate both the learning environment students in which the students took part in the course and the resulting students' written texts that are at core of this analysis.

Chapter 4

Methodology

4.1 Course Design

The units of analysis for this work come from the written submissions of students in three different language classes. All students in this sample were enrolled in distance education courses at the University of Waterloo. The three courses were German 101 (Elementary German I), German 102 (Elementary German II) and German 201 (Intermediate German I) offered by the Department of Germanic and Slavic Studies.

All components of the course were accessible through UW-ACE, a virtual learning environment offered by the University of Waterloo. All three courses are structured around the second edition of the textbook *Vorsprung: An Introduction to the German Language and Culture for Communication* (Lovik, Guy and Chavez, 2007) and focus on different chapters of the book. All of the courses were taught by the same instructor but several teaching assistants were responsible for marking the students' submissions. If they did not understand an assignment or the marking scheme, students could seek help from both the instructor and the teaching assistants assigned to the student.

The marking scheme was the same across all courses. All three courses required students to read particular textbook pages during specific times. Students in all three courses had to submit six tasks for grading. Topics of these tasks include but are not limited to descriptions of the learners' friends, families, living arrangements, hometowns, countries and hobbies. The tasks made up 50% of the final grade for each course. The submission of each task called for the successful completion of a practice quiz. Students had to acquire a minimum of 70% on each quiz in order to proceed. Unlimited trials were allowed for successful completion of the quiz.

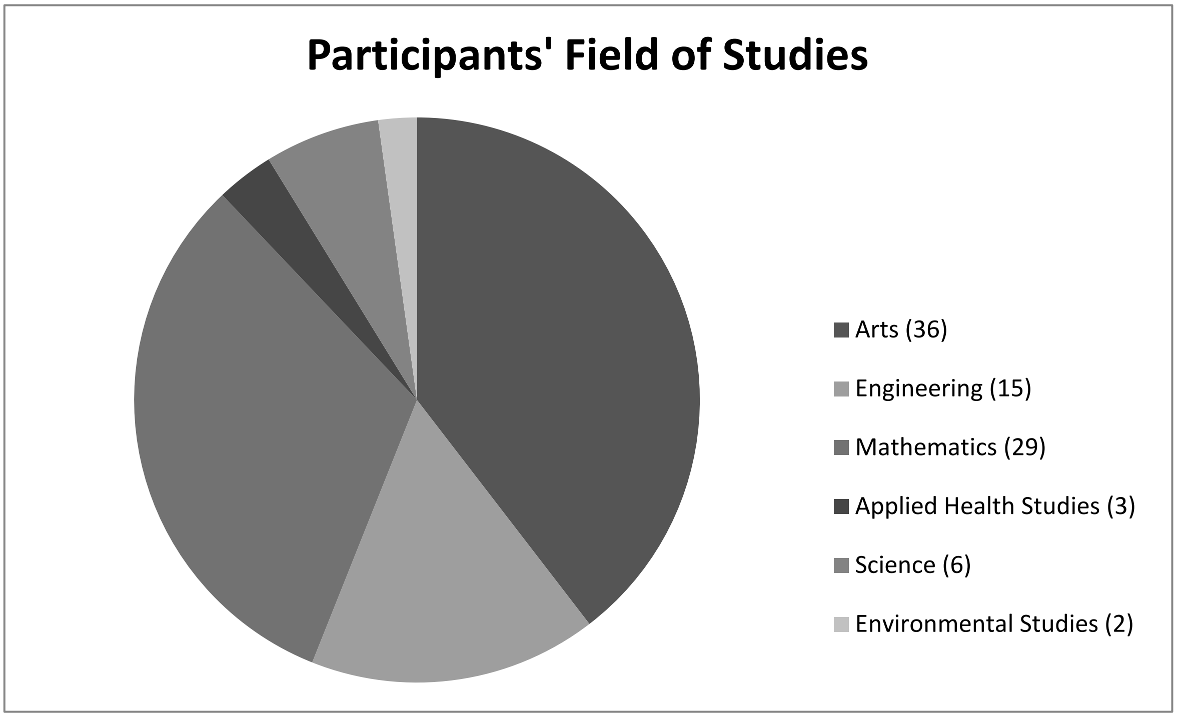
Students in all three courses were asked to participate in discussion boards about the course content. Participation in these discussions counted for their participation grade and accounted for 10% of the students' grades. A topic was always assigned for these discussion boards and students were encouraged not only to post their submissions regularly but also to respond to the submissions of other students. These responses could include but were not limited to corrections on ill-formed submissions of peers, encouragement and praise for correct submissions or follow-up questions to facilitate further discussion amongst the learners. Students were encouraged to complete these submissions in German. This language requirement was largely respected. If part of a comment was written in English, the English text was excluded from the analysis. The extent to which students wrote on the discussion boards varied greatly within time intervals, courses and terms. Lastly, students in all three courses completed a final, 2.5 hour proctored, course specific, written exam for their final course credit. The final exam accounted for the remaining 40% of the course grade.

The content and grading scheme of the courses differed little from the on-campus versions of the above mentioned courses. A study carried out by Schulze, Liebscher and Su (2007) set out to examine if any differences were present between the performance of students taking these courses on campus and students completing the courses online. In an effort to increase reliability, the authors compared student performance on the midterm and the final examination as these two testing units are graded collectively by the same group of teaching assistants. The authors found no significant difference between the performance of on-campus and distance education students, concluding the distance education course is comparable in outcome as measured by student performance to the same course offered on campus. Given that only minor changes were made to the course design since the research study was carried out, one can conclude that Schulze, Liebscher and Su's findings from 2007 can be transferred to the students' performance in 2008 and 2009.

4.2. Participants

The participants in this study were 91 students at the University of Waterloo. All but three participants were undergraduate students. The sample consisted of 39 male students and 52 female students. The sample included 9 first-year students, 25 second-year students, 28 third-year students and 26 fourth-year students. Information about the participants' fields of study is available in Figure I: Participants' fields of studies.

Figure I: Participants' fields of studies



The above information allows for the conclusion that the sample used in this study is a very diverse one and is representative of the overall student body in German classes at the University of Waterloo.

4.3. Materials

Prior to commencing the analysis, I relied on two tools to prepare all units of this work: *Björn's Online Spell Checker* and the *Morphy lemmatizer* for the German language. The analysis itself relied on the use of subject specific word lists discussed above, two pieces of software (*Range* and *Frequency)* and the students' texts. Each unit will be discussed below.

4.3.1. *Björn's Online Spell Checker*

This spell checker is available as a free online tool for personal or research purposes.[1] The user can choose to have the text corrected for the old or new German spelling conventions and for this work the latter was selected. All word lists and students' submissions were analyzed by this spell checker and the incorrectly spelled words were corrected. In instances where the spell checker offered more than one option for a misspelled item, the more correct entry as judged by the researcher was selected.

The reason for spellchecking all vocabulary lists and submissions will be discussed in chapter 4, section 4.4.1.

4.3.2. *Morphy*

Created by Wolfgang Lezius[2] – a computer linguist at the University of Paderborn – *Morphy* is a lemmatizer for the German language and is available for research purposes without charge. It can be downloaded from Lezius' web site as a complete package which includes the documentation and the lexicon. *Morphy's* lexicon comprises 324,000 word forms based on 50,500 stems (Lezius, Rapp and Wettler, 1998). The reason for lemmatizing all of the materials used in the analysis will be presented in chapter 4, section 4.4.1.

[1] http://www.j3e.de/cgi-bin/spellchecker

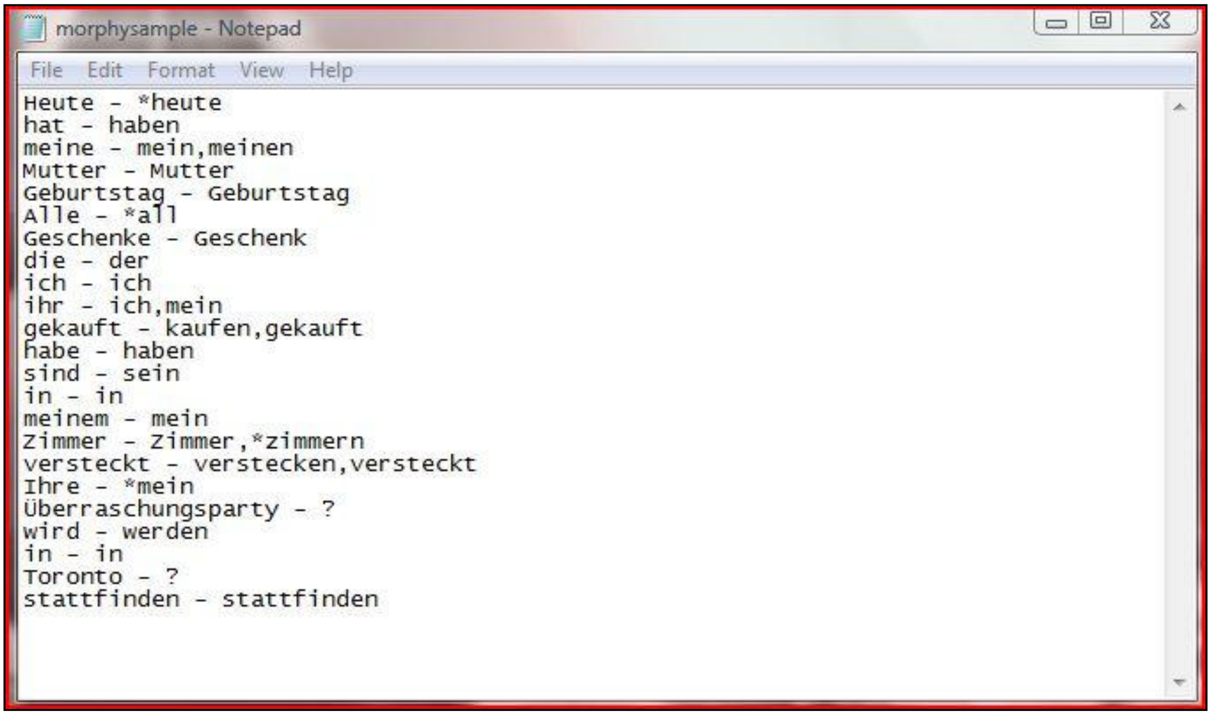
[2] http://www.wolfganglezius.de/doku.php

This software offers several tools frequently used in research on natural language processes but for the purpose of this analysis only features pertaining to the morphological analysis options will be discussed.

Once the desired file is analyzed, *Morphy* informs the user about the length of time the analysis took, the number of words that were analyzed and the number of items the program did not recognize. A sample output of *Morphy* is available in Figure II: *Morphy* Sample Output. The analysis was carried out on the following sample text generated for illustrative purposes:

"*Heute hat meine Mutter Geburtstag. Alle Geschenke, die ich ihr gekauft habe, sind in meinem Zimmer versteckt. Ihre Überraschungsparty wird in Toronto stattfinden*." [Today is my mother's birthday. All of the gifts that I purchased for her are hidden in my room. Her surprise party will take place in Toronto]

Figure II: *Morphy* Sample Output



```
Heute - *heute
hat - haben
meine - mein,meinen
Mutter - Mutter
Geburtstag - Geburtstag
Alle - *all
Geschenke - Geschenk
die - der
ich - ich
ihr - ich,mein
gekauft - kaufen,gekauft
habe - haben
sind - sein
in - in
meinem - mein
Zimmer - Zimmer,*zimmern
versteckt - verstecken,versteckt
Ihre - *mein
Überraschungsparty - ?
wird - werden
in - in
Toronto - ?
stattfinden - stattfinden
```

Several features of this analysis must be pointed out. Since *Morphy* outputs the results by placing one item per line, all further files analyzed by *Morphy* reflect this pattern as well.

The left side of the image displays the entered text, while all entries to the right of the hyphen - are the results of the morphological analysis. First, all derivations of a verb are lemmatized to the infinitive form as evidenced in "*hat* → *haben*" [has → to have] and "*gekauft*→ *kaufen*" [purchased → to purchase]. In addition to the infinitive, the perfect form is also offered. For the purpose of this analysis, the lemmatized results were edited to include only the infinitive verb form, and all perfect forms were deleted.

The presence of an asterisk * indicates that the lemmatized item can also appear with a capital letter at the start of the item, such as in "heute/Heute" [today]. Instances where the item to be lemmatized is followed by a hyphen and a question mark indicate that *Morphy* was unfamiliar with the word and could therefore not analyze it. An approach to deal with such instances was standardized for this work. The absence of a lemma for the item "Toronto" was not problematic as proper nouns were excluded from all students' texts. The reason for this is discussed in Chapter 4, section 4.4.1. Despite the absence of a lemma for the item "*Überraschungsparty*" [surprise party], an item like this was included as it was spelled correctly and was semantically accurate. More specific guidelines for the inclusion rate of items are offered in a subsequent section.

The default gender for a lemma with *Morphy* is masculine. As such, all articles are lemmatized as "der"; all possessive pronouns are lemmatized as "*mein*" [my] and so forth. The exception to the rule pertains to personal pronouns, all of which were lemmatized as "*ich*" [I]. Implications of this will be discussed in connection with *Range*, the software that analyzed the students' submissions with reference to the aforementioned vocabulary lists.

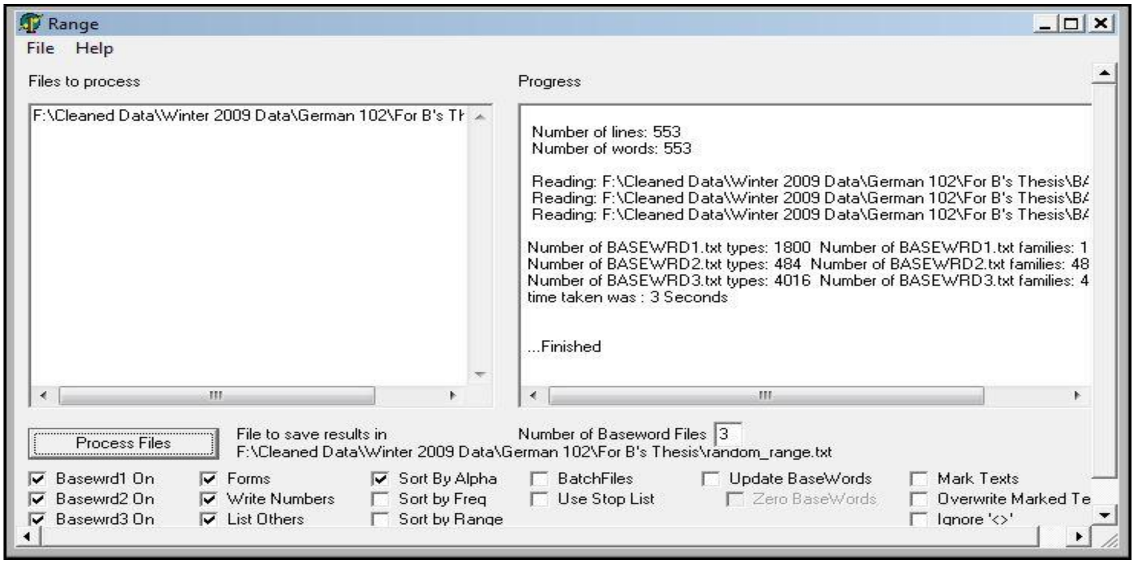
As shown in the above sample analysis, *Morphy* often offered two lemmas for a given item. In instances where this occurred, one of the suggested lemmas was persistently incorrect and could be excluded because of the context. An example of such would be the lemma for "*Zimmer*" [room], which once analyzed is presented as "*Zimmer*" or "*zimmern*" [to carpenter]. When this occurred, a description was made based on the content of the text which of the two lemmas was more accurate. From the sample given above, it is clear that the text makes reference to a room and not carpentry and therefore, the lemma "*Zimmer*" was selected. An additional reason for two lemmatized results is the frequency of homonyms in the German language and *Morphy's* necessary disregard for capitalization. An example of this would be the above presented item "*ihr*" which once lemmatized appears as "*ich*" or "*mein*" as it could refer to second person plural pronoun, the third person singular possessive pronoun or third person singular dative form. The decision as to which lemma to keep in the analysis was based on the content of the submission. In the above example, "*ihr*" refers to the third person singular personal pronoun and the lemma "*mein*" would have been excluded.

4.3.3. *Range*

Developed by Paul Nation for research purposes, the *Range* software is available free on the author's website (Nation, 2005). Range is utilized without installation. Numerous features are available to users but only features pertaining to the present work will be discussed here.

Range can be used to analyze up to 32 different texts at the same time. For each word in the sample text, it can provide a distribution figure indicating the number of texts in which a given word occurs. An image depicting a completed analysis by *Range* can be seen in Figure III: Range.

Figure III: Range



F:\Cleaned Data\Winter 2009 Data\German 102\For B's Tł

```
Number of lines: 553
Number of words: 553

Reading: F:\Cleaned Data\Winter 2009 Data\German 102\For B's Thesis\B/
Reading: F:\Cleaned Data\Winter 2009 Data\German 102\For B's Thesis\B/
Reading: F:\Cleaned Data\Winter 2009 Data\German 102\For B's Thesis\B/

Number of BASEWRD1.txt types: 1800  Number of BASEWRD1.txt families: 1
Number of BASEWRD2.txt types: 484  Number of BASEWRD2.txt families: 48
Number of BASEWRD3.txt types: 4016  Number of BASEWRD3.txt families: 4
time taken was : 3 Seconds


...Finished
```

File to save results in F:\Cleaned Data\Winter 2009 Data\German 102\For B's Thesis\random_range.txt

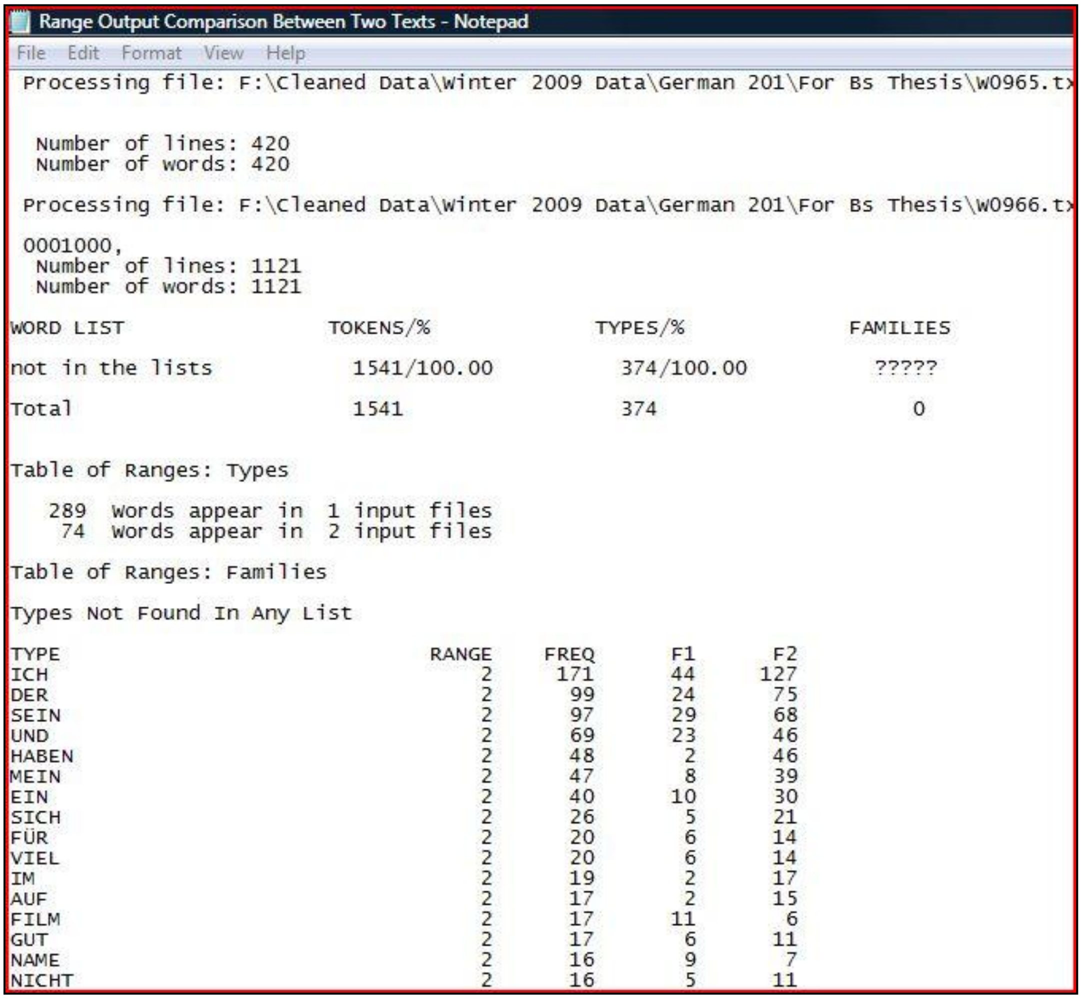
Number of Baseword Files 3

☑ Basewrd1 On ☑ Forms ☑ Sort By Alpha □ BatchFiles □ Update BaseWords □ Mark Texts
☑ Basewrd2 On ☑ Write Numbers □ Sort by Freq □ Use Stop List □ Zero BaseWords □ Overwrite Marked Te
☑ Basewrd3 On ☑ List Others □ Sort by Range □ Ignore '<>'

The left window shows the files that the software has analyzed, while the right window displays the analysis as it progresses. The check boxes below the windows allow the user to customize and refine the analysis.

The primary functions that attracted me to this software are *Range's* ability to carry out an examination of the similarities and differences between texts. A sample output of a comparison between two texts is depicted in Figure IV: Range Output: Comparison between two texts.

Figure IV: Range Output: Comparison between two texts


Range Output Comparison Between Two Texts - Notepad
File Edit Format View Help

```
 Processing file: F:\Cleaned Data\Winter 2009 Data\German 201\For Bs Thesis\w0965.tx


  Number of lines: 420
  Number of words: 420

 Processing file: F:\Cleaned Data\Winter 2009 Data\German 201\For Bs Thesis\w0966.tx

 0001000,
  Number of lines: 1121
  Number of words: 1121

WORD LIST                 TOKENS/%            TYPES/%             FAMILIES

not in the lists            1541/100.00         374/100.00          ?????

Total                       1541                374                     0


Table of Ranges: Types

   289  Words appear in  1 input files
    74  Words appear in  2 input files

Table of Ranges: Families

Types Not Found In Any List

TYPE                              RANGE     FREQ      F1      F2
ICH                                   2      171      44     127
DER                                   2       99      24      75
SEIN                                  2       97      29      68
UND                                   2       69      23      46
HABEN                                 2       48       2      46
MEIN                                  2       47       8      39
EIN                                   2       40      10      30
SICH                                  2       26       5      21
FÜR                                   2       20       6      14
VIEL                                  2       20       6      14
IM                                    2       19       2      17
AUF                                   2       17       2      15
FILM                                  2       17      11       6
GUT                                   2       17       6      11
NAME                                  2       16       9       7
NICHT                                 2       16       5      11
```

As shown in Figure IV, *Range* offers a count of the number of words and lines in each text, the type token ratio for both combined texts and qualitative information about the occurrence of each word. Furthermore, because all files were lemmatized by *Morphy*, the number of lines and words are the same for every text, with one word per line.

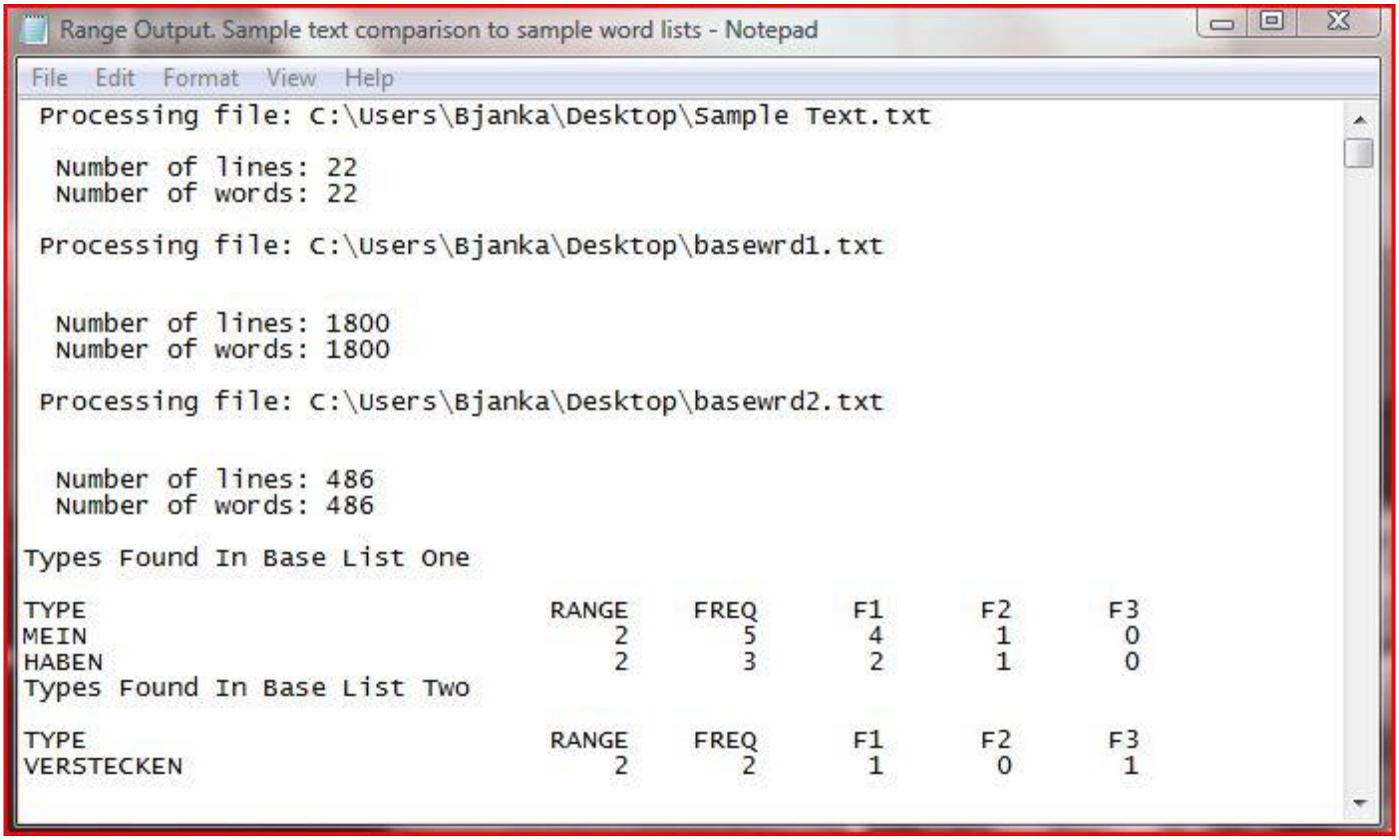
The *Range* output is read as follows: the item “*NICHT*” [not] occurs in both texts a total of 16 times. This item occurs in the first text five times and eleven times in the second text.

The implications of the preceding discussion of the lemmatization process must be discussed. The sample outputs from *Range* must be interpreted in light of the discussion of *Morphy’s* default settings. In Figure II the item “*ich*” [I] appears 171 times. This number represents the occurrence of all personal pronouns, regardless of gender or number. Similarly, the item “*mein*” represents all possessive pronouns and not just first person singular.

Range can also find tell the user how many words from a word list are in a text, create word lists based on frequency and range, and discover shared and unique vocabulary in different pieces of writing. *Range* can compare a text against a vocabulary list to see what words in the text are and are not in the list and to see what percentage of the words in the text are covered by the list. A sample output of *Range’s* comparison of the sample text to two word lists is depicted in Figure V: Range Output: Sample text comparison to sample word lists.

As shown, all units involved in the analysis are described in terms of length. The first text is the list to be analyzed while the two subsequent units are the list that will be used for comparison. Figure V can be read as follows. The lemma “*mein*” occurs in two locations (RANGE 2). In this case, the first location is the text that was analyzed (F1) and the second location is the first word list against which the text was compared (F2). In the text that was analyzed, “*mein*” occurs for a total of five times (FREQ 5). “*Mein*” does not occur in the second vocabulary list at all (F3 0). Similarly, the lemma "*verstecken*" [to hide] also occurs in two locations that were shown here (RANGE 2). This item occurs once in the text that was analyzed (F1) and once in the second vocabulary list that was used (F3).

Figure V: Range Output: Sample text comparison to sample word lists


Range Output. Sample text comparison to sample word lists - Notepad
File Edit Format View Help

```
 Processing file: C:\Users\Bjanka\Desktop\Sample Text.txt

  Number of lines: 22
  Number of words: 22

 Processing file: C:\Users\Bjanka\Desktop\basewrd1.txt


  Number of lines: 1800
  Number of words: 1800

 Processing file: C:\Users\Bjanka\Desktop\basewrd2.txt


  Number of lines: 486
  Number of words: 486

Types Found In Base List One

TYPE                             RANGE    FREQ      F1      F2      F3
MEIN                                 2       5       4       1       0
HABEN                                2       3       2       1       0
Types Found In Base List Two

TYPE                             RANGE    FREQ      F1      F2      F3
VERSTECKEN                           2       2       1       0       1
```

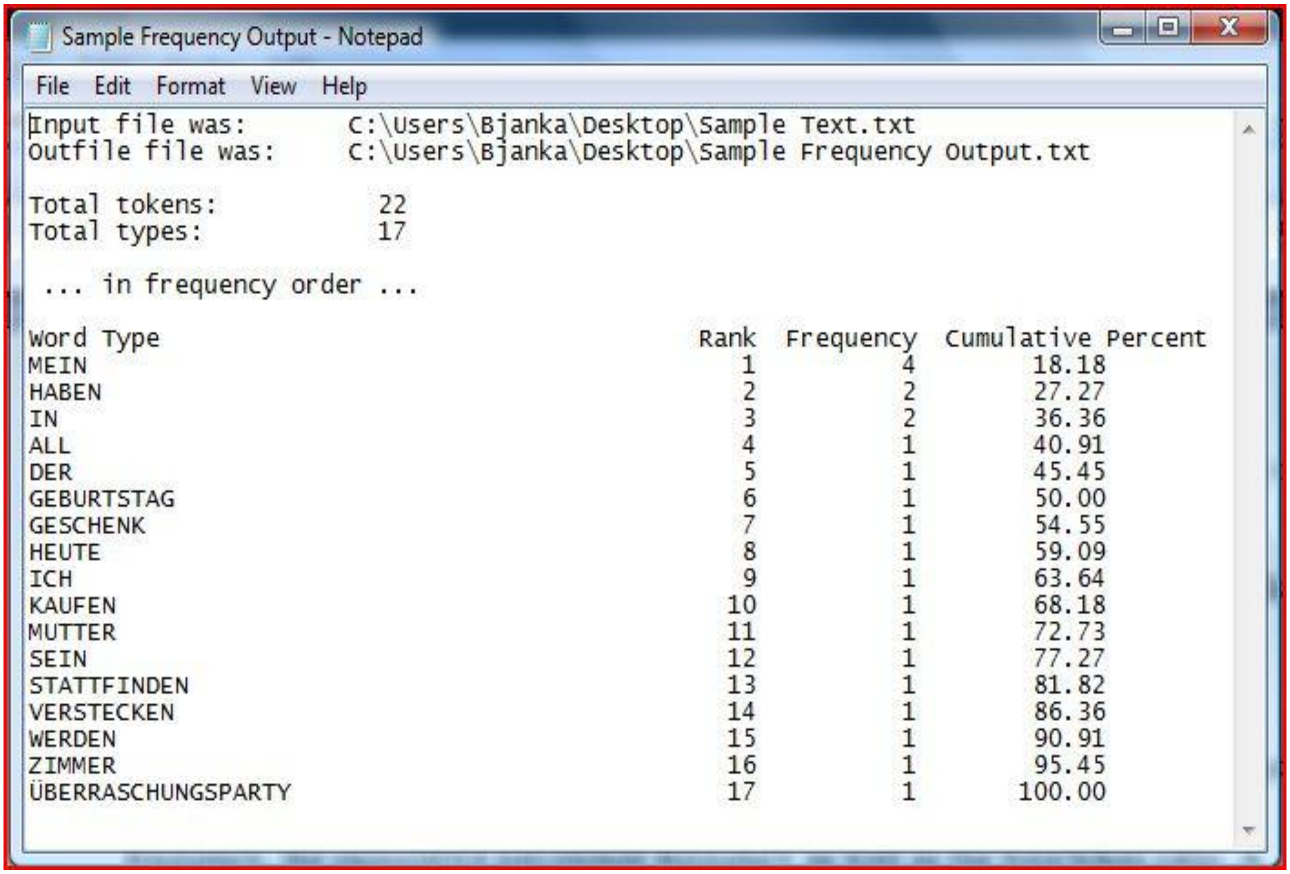
Range stipulates that the user labels all the word lists that will be used for comparison under the names basewrd1, basewrd2 and so forth. A limitless number of lists could be used for comparison as long as the naming stipulation is followed.

For the purpose of this analysis, three vocabulary lists were used for comparison but were always used in the same order. The *Vorsprung* vocabulary list was consistently the primary list (basewrd1) while the Subtitles and *PDW* vocabulary lists were always the second and third lists (basewrd2 and basewrd3) respectively. The vocabulary lists were placed in that order because I assumed that the majority of words used will be from the *Vorsprung* list. The *PDW* vocabulary list was placed after the Subtitles list because of the assumption that students would use words from this list the least because of this list was based on such a high register of German.

4.3.4. *Frequency*

Frequency is the final software used in this analysis and is also available for research purpose at no cost (Nation, 2005). The program *Frequency* creates frequency lists of all the words in a text. It can only analyze one text at a time and can create an output file as an alphabetical or frequency ordered list. It offers the rank order of all words, their raw frequency, the cumulative percentage frequency as well as the type token ratio. A sample output of *Frequency's* analysis of the same sample text mentioned previously can be viewed in Figure VI: Sample *Frequency* Output

Figure VI: Sample *Frequency* Output



```
Input file was:      C:\Users\Bjanka\Desktop\Sample Text.txt
Outfile file was:    C:\Users\Bjanka\Desktop\Sample Frequency Output.txt

Total tokens:          22
Total types:           17

 ... in frequency order ...

Word Type                                    Rank  Frequency  Cumulative Percent
MEIN                                            1          4        18.18
HABEN                                           2          2        27.27
IN                                              3          2        36.36
ALL                                             4          1        40.91
DER                                             5          1        45.45
GEBURTSTAG                                      6          1        50.00
GESCHENK                                        7          1        54.55
HEUTE                                           8          1        59.09
ICH                                             9          1        63.64
KAUFEN                                         10          1        68.18
MUTTER                                         11          1        72.73
SEIN                                           12          1        77.27
STATTFINDEN                                    13          1        81.82
VERSTECKEN                                     14          1        86.36
WERDEN                                         15          1        90.91
ZIMMER                                         16          1        95.45
ÜBERRASCHUNGSPARTY                             17          1       100.00
```

The output is read as follows: The word “*mein*” is the most frequently occurring word. It appears a total of four times and takes up 18.18% of the text. The word “*haben*” [to have] is the second most frequent word, occurs twice. The second item makes up 9.09% of the text or 27.27% of the text when considered with the most frequent item occurring right before it. The items that occur only once are alphabetized and share equal intervals to the preceding and subsequent words.

Having introduced the software used in the project, a discussion about how the students' submissions were obtained and prepared for analysis, will be presented.

4.4. Procedure

Upon logging in to UW-ACE for the first time, each student was presented with a consent form. This document explained that the student had the option of participating in a study carried out by researchers within the department. The consent form outlined that participation did not call for any additional work but rather required the students to give permission to the researchers to use their submissions for research purposes.

Anonymity was stressed and students were assured that they could withdraw from the study at any point without suffering academic consequences. No reimbursement was offered to the student in exchange for their participation. The students' submissions were downloaded three times during the semester and only the submissions of students who agreed to participate were included in the analysis.

4.4.1. Preparation of all materials used in the study

All of the students' submissions and the frequency lists were modified in the same manner. Misspelled words in the students' submissions and/or word lists were identified with the help of *Björn's Online Spell Checker*. Corrections suggested by the software were implemented.

The purpose of this step was two-fold. The spell checker firstly allowed for the standardization of the spelling amongst all documents used in this study. This was carried out so that items in the students' submissions could be compared to any vocabulary lists without being influenced by differences in spelling conventions. Secondly, the spell checker allowed for the correction of incorrectly spelled items as they would not be recognized by either *Morphy Lemmatizer* or *Range*. Misspelled items would be recognized by *Frequency* as this software does not rely on any external sources to complete its analysis.

This procedure of spellchecking all documents was also suggested by Laufer and Nation (1995), who write that spelling errors could greatly skew the resulting analyses. This step does pose a limitation for the present work as I cannot claim that the units analyzed reflect the true efforts of learners as no distinction is made between learners whose work was generally spelled correctly and those whose spelling mistakes were abundant. I considered this a small compromise to make in order to obtain more robust results.

Furthermore, all of the students' texts were modified according to the instructions outlined by previous researchers using the Lexical Frequency Profile. Laufer and Nation (1995) recommend a minimum text length of 300 words for analysis. Texts shorter than that were judged to be "unstable" by the authors as results were skewed due to a high type token ratio.

This suggestion was easily adopted for the present work. All submissions and the discussions of the students were separated from each other and saved in individual files. Thus any one file contained all the work of one student throughout the term. Once combined, all of these written texts were usually well over 300 words. Exceptions did exist of course and submissions shorter than the required 300 words were excluded from the analysis. This occurred for a total of 22 students, leaving a total sample size of 69 students. I speculate that instances where the combined submissions by one student were under 300 words are cases in which the student dropped the course and did not complete all of the assigned tasks.

Laufer and Nation further suggest that clearly incorrect words should be removed from the analysis as they cannot be considered to be known by the learner. The term "clearly incorrect" is for the researcher to define based on a given study. The items deleted from the data at hand can be grouped into two categories: instances where the item is not recognized or understood and instances where an item is semantically incorrect. These were excluded through the use of the spell checker and by being proofread by the researcher.

An example of an item that is misspelled beyond recognition would be: "*Er hat dir gut gefallen wenn wir sind gegenseitig **ähgnel***" (Student W0959). One could argue that the learner attempted to write "*ähnlich*" meaning similar instead of "*änghel*" but this would be speculative at best. Even if this is in fact the word the learner wanted to use, one cannot argue that the learner demonstrates knowledge of the German word for "similar". For this reason, the items and other similar items were excluded.

An example of a semantically incorrect example is: "*Sie hat von Irland **bewegt**"* (Student W0960). While the sentence is comprehensible to somebody who speaks both English and German, "*bewegen*" is the incorrect verb for this context, and its correct use or the correct verb for this sentence is presumably not known by the learner. For this reason, this item and instances similar were excluded from the submissions.

Laufer and Nation suggest that misspelled words be corrected and then included as in the example "*Meine Eltern denken ich soll Museen **besuhen***. (Student W0960*)"*. In this example, “*besuchen*” is the correct spelling. The verb is correct and the one can therefore assume that the student is familiar with its meaning and use.

One more frequently corrected spelling mistake must be mentioned. A large proportion of students in German 101 (Elementary German I) often used the lexeme “B” in favor of the German lexeme “ß”. To a layperson, the sentence “*die Studentin heiBt Anna*” might be as incomprehensible as it was to *Björn’s Online Spell Checker.* Once considered in light of the fact that it was authored by a beginner who is just learning the orthographic rules of the German language and does not yet know how to represent all the new graphemes on the computer, the intended meaning of “*die Studentin heißt Anna*” [the student’s name is Anna] becomes apparent.

Laufer and Nation further suggest that any incorrect derivatives if they occur should not be considered as errors as they belong to the same word family. On this basis, ill-formed derivatives such as incorrect verb conjugations were included in the analysis. Examples of such include sentences like “*Ich **woll**” (*Student W0960). Although a grammatical error, the word “*woll*” was treated essentially like a spelling error and corrected to the right conjugation “*will*”.

Furthermore, Laufer and Nation also call for the removal of proper nouns from any texts to be analyzed. The rationale for this suggestion is the fact that proper nouns are often of low frequency and their presence in such an analysis would skew the results. Therefore, it is understandable why the absence of a lemma for the item “Toronto” does not create a problem for this analysis as “Toronto” would have been removed from the initial analysis. Frequently occurring proper nouns deleted from the students’ submissions for this work include but are not limited to "*Tim Horton's*", "*Maple Leafs*", "*Oktoberfest*" as well as the names of any national holidays, persons or geographical locations.

In addition to the above mentioned exclusions, all numeric values present in students' submissions were also deleted. Because there was no agreement in the way students used numbers in their submissions (eg. "*Er ist sechs Jahre alt*" vs. "*Er ist 6 Jahre alt*") [he is six years old]; it was deemed appropriate to exclude numbers from the analysis.

Lastly, because *Range* does not recognize the German grapheme "ß", its occurrences in all of the submissions and vocabulary lists were replaced with two hyphens --. This was done to ensure that *Range* successfully includes all word in the analysis. Previous trials showed that *Range* ignored the grapheme "ß" and considered the preceding and following letters as two lexical entries. Thus "*heißen*" would be broken down into "*hei*" and "*en*" and neither item would be recognized as it is not featured in any vocabulary list. The replacement of two hyphens was selected as such a combination did not occur in any students' texts or lists. Once all instances of "ß" were replaced with two hyphens, *Range* easily recognized the entered items and could easily reference them to the appropriate vocabulary lists.

The next step taken to prepare all items necessary for the analysis was the lemmatization of all vocabulary lists and submissions. Doing so had both advantages and disadvantages, but I felt that the former greatly outweighed the latter.

The primary advantage associated with this step is that the lemmatization process greatly reduced the efforts invested in creating the vocabulary lists used for comparison. Unlike English, the German language is highly inflectional. This inflection is something

Laufer and Nation did not have to consider in their original study. The many prefixes, suffixes and markings of gender, case and tense are just a few things that can greatly change the form of a word without any change to the semantics. Without lemmatization, every single one of these forms would have to be included in the vocabulary lists used for comparison.

This predicament is perhaps best illustrated through an example: The *Vorsprung* vocabulary list contains the form "*groß"* [big]. Two students could hypothetically use this word "*groß*" but to describe two items of different gender such as "*mein großes Buch*" [my big book] and "*mein großer Hund*" [my big dog]. Without lemmatization, neither item would be traced back to the *Vorsprung* list because all three forms of *"groß"* differ. Despite the fact that they convey the same meaning, because these items are not identical in form, *Range* would consider "*großes*" and "*großer*" as not only two different words, but also completely unrelated to "*groß"*. Thus, *Range* would be unable to trace these two forms back to the *Vorsprung* list but would rather list them as "not appearing on any list". Items listed as not appearing in any list are items of low frequency. This would in turn increase the number of words that were not accounted for through any vocabulary list, thus skewing the results and giving the false indication that the students used highly infrequent words. This is of course absurd as *Vorsprung* introduces the learner to*"groß"* in the first chapter as well as to the comparative or superlative form of the adjective in German 101.

It is important to note that the above described example and *Range's* approach towards word form is not a fault of *Range* as a piece of software, but rather a feature of the German language that had to be accommodated. The inclusion of every form of all items in the three vocabulary list is counterintuitive. Doing so would be time-consuming and would greatly increase the size of every list without increasing the benefit derived from the presence of these word forms. This is but one reason why the lemmatization process was deemed necessary for this work.

I acknowledge that this lemmatization process deprives the students' written submissions of all grammatical features. Once lemmatized, the students' submissions are virtually indistinguishable with reference to grammatical complexity and accuracy despite the fact that the students clearly differ in these areas as evidenced by the fact that they are enrolled in different level courses. Therefore, no conclusions can be drawn about students who use "*groß*" in its base, comparative and superlative form. This is fortunately not a drawback for this project as vocabulary use and not grammatical accuracy are being analyzed. Put differently, by stripping the students' submissions of all grammatical components, the lemmatization process allowed for a more detailed analysis of the topic of interest and reduced the possibility that any grammatical components could skew the results.

The disadvantage to the lemmatization process is the astounding amount of time necessary to lemmatize all vocabulary lists and students' submissions. The reader can refer back to Figure II and the discussion pertaining to this figure for a reminder on the steps needed to clean each lemmatized file.

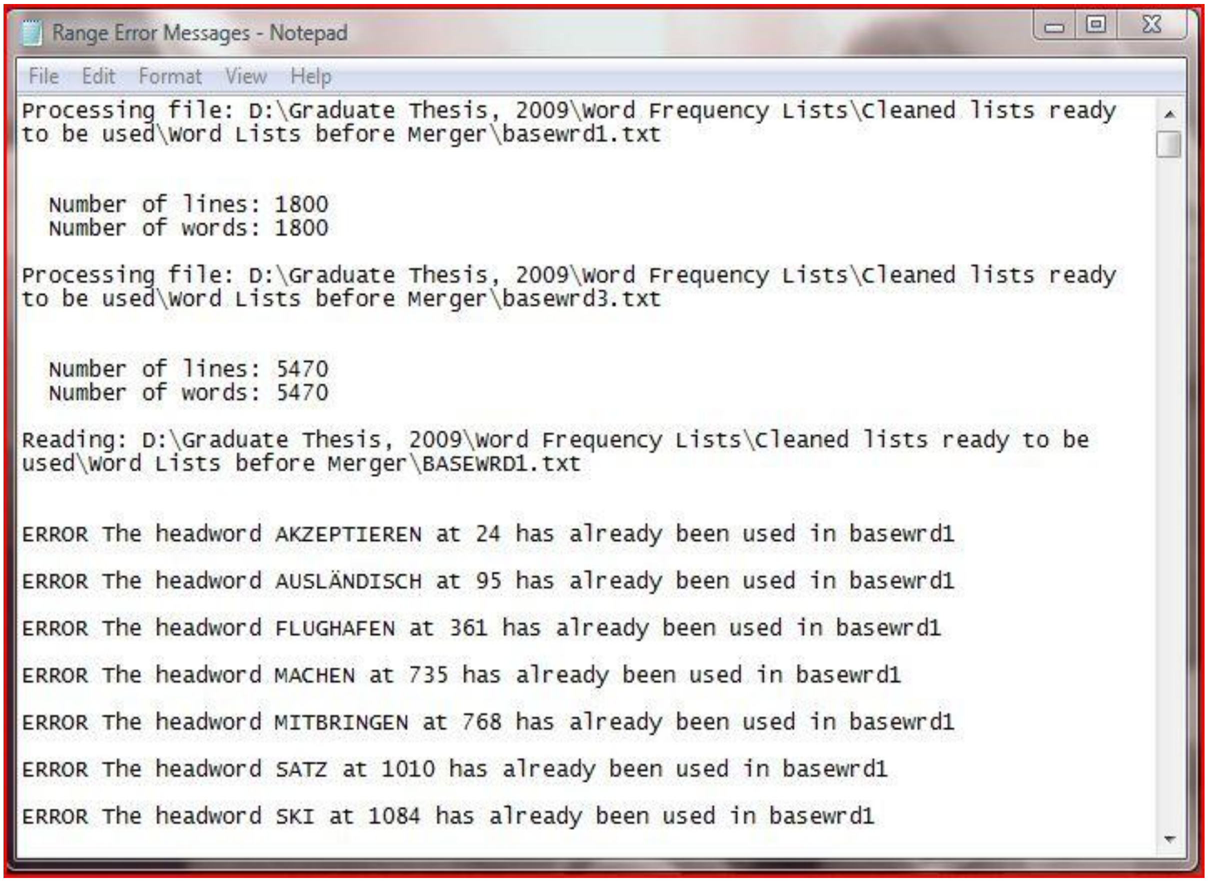
One final step was needed to successfully integrate the word lists into this project. As mentioned, all word lists were analyzed by *Morphy*, providing lemmas of the individual items found in the list. The resulting list included several repeating lemmas.

This is better understood through an example: Items like "*groß*", "*große*", "*größte*" would all be lemmatized into "*groß*", and the lemma "groß" would appear three times within one list. Because *Range* does not allow repetition within any one vocabulary list or between the vocabulary lists against which a text is compared, all repeated lemmas had to be removed. This step resulted in significantly shorter lists. To illustrate, the *Vorsprung* list originally contained 2,534 items, the Subtitles list originally contained 2,000 items and the *PDW* list consisted of 10, 000 items. After being lemmatized by *Morphy* and after the exclusion of repeating lemmas, the sizes of the lists were changed to 1,800, 1,091 and 5,470 items respectively.

Following the deletion of repeated lemmas, the three lists (*Vorsprung* lemmas, Subtitle lemmas and *PDW* lemma) were standardized so that no one lemma appeared in more than one list. This was done to configure the lists for *Range*, as the software does not allow for any one item to appear in more than one list. The analysis is not executed if repeated lemmas are found within either list or between lists. This was accomplished by running all the three lists against each other. *Range* automatically singles out all repeating lemmas as errors (see Figure VII: Repeated Items in Word Lists for illustration).

For this analysis, all items labeled as errors were manually deleted from the lists. Thus, all *Vorsprung* lemmas that also appeared in the Subtitle list and the *PDW* list were deleted. This group consisted of 606 lemmas.

Figure VII: Repeated Items in Word Lists



```
Processing file: D:\Graduate Thesis, 2009\Word Frequency Lists\Cleaned lists ready
to be used\Word Lists before Merger\basewrd1.txt


  Number of lines: 1800
  Number of words: 1800

Processing file: D:\Graduate Thesis, 2009\Word Frequency Lists\Cleaned lists ready
to be used\Word Lists before Merger\basewrd3.txt


  Number of lines: 5470
  Number of words: 5470

Reading: D:\Graduate Thesis, 2009\Word Frequency Lists\Cleaned lists ready to be
used\Word Lists before Merger\BASEWRD1.txt


ERROR The headword AKZEPTIEREN at 24 has already been used in basewrd1

ERROR The headword AUSLÄNDISCH at 95 has already been used in basewrd1

ERROR The headword FLUGHAFEN at 361 has already been used in basewrd1

ERROR The headword MACHEN at 735 has already been used in basewrd1

ERROR The headword MITBRINGEN at 768 has already been used in basewrd1

ERROR The headword SATZ at 1010 has already been used in basewrd1

ERROR The headword SKI at 1084 has already been used in basewrd1
```

All remaining Subtitle lemmas that also occurred in the *PDW* list were deleted from the latter list. This group consisted of 909 lemmas. Once this was carried out, the *Vorsprung* list remained at the same size of 1,800 lemmas, the Subtitles vocabulary list decreased to 484 items and the *PDW* list shrank to 4016 words.

This step created three distinct vocabulary lists. This was of great benefit for the present work. Firstly, any one student text could be run against all three lists simultaneously, which substantially reduced the amount of time spent on analysis. Secondly, when one input file is compared against all three vocabulary lists at once, the resulting output is easier to read and any one word is immediately classified as belonging to either one of the vocabulary lists or is labeled as not being found in any of the three vocabulary lists. Lastly, the large overlap between the three lists signifies that items in the *Vorsprung* and Subtitles vocabulary lists are of high frequency as they also appear in the *PDW* vocabulary list.

The significance of removing lemmas between the lists must be discussed at this point. Little information was available online about the Subtitles list. I believe that the quality of the list has been validated given that so many of the items from this list overlapped with two already established lists. In addition, given that the overlapping lemmas have been removed from lists, one must keep in mind that should an item be classified as belonging to the *Vorsprung* list, this is not to mean that it is found solely in this source, but could have as well been included in the original version of the other two lists.

Since all relevant tools and their preparation and utilization for this work have been introduced, I will now discuss how the data were analyzed. Firstly, each of the three datasets was analyzed in isolation. Descriptive statistics will be presented first. These will include discussions of average text length and type token ratio.

The type token ratio is a frequently used measure of lexical diversity. This measure is frequently criticized as it is heavily dependent on text length, thus making it less reliable. To avoid this shortcoming, the present study will rely on Carroll's type token ratio, an approach that controls for text length. Instead of simply dividing the number of types by the number of tokens as the original formula demands, this approach calls for the division of the number of types (t) by the square root of twice the number of words (w) (Schulze, Wood and Pokorny, unpublished manuscript). The final formula is represented as shown: $t/\sqrt{(2w)}$. The higher the resulting number is, the more lexical variety the sample includes.

Attention will then be devoted to the lemmas common to a number of submissions for each individual sample. For each sample, the words that occur in all submissions will be discussed first. Next, the findings for each sample will be grouped into percentiles and analyses will be carried out for each percentile group by comparing the percentiles to the three vocabulary lists. The analysis of each individual dataset will end with the mean number of words used from each vocabulary list and a series of t-tests that will compare these scores.

Comparisons will then be made between the three samples. The vocabulary common to all three samples will be discussed. Before concluding, a series of t-tests will be presented and their significance will be discussed. These were carried out to determine if the differences found between the datasets are statistically significant. Interpretations and conclusions will be offered in the closing stages.

Chapter 5

Results

5.1. Dataset I: Analysis of the Submissions for German 201:

The submissions of all eleven German 201 students were analyzed with reference to each other. This was done to extract some descriptive statistics and the common vocabulary between the texts of all eleven students.

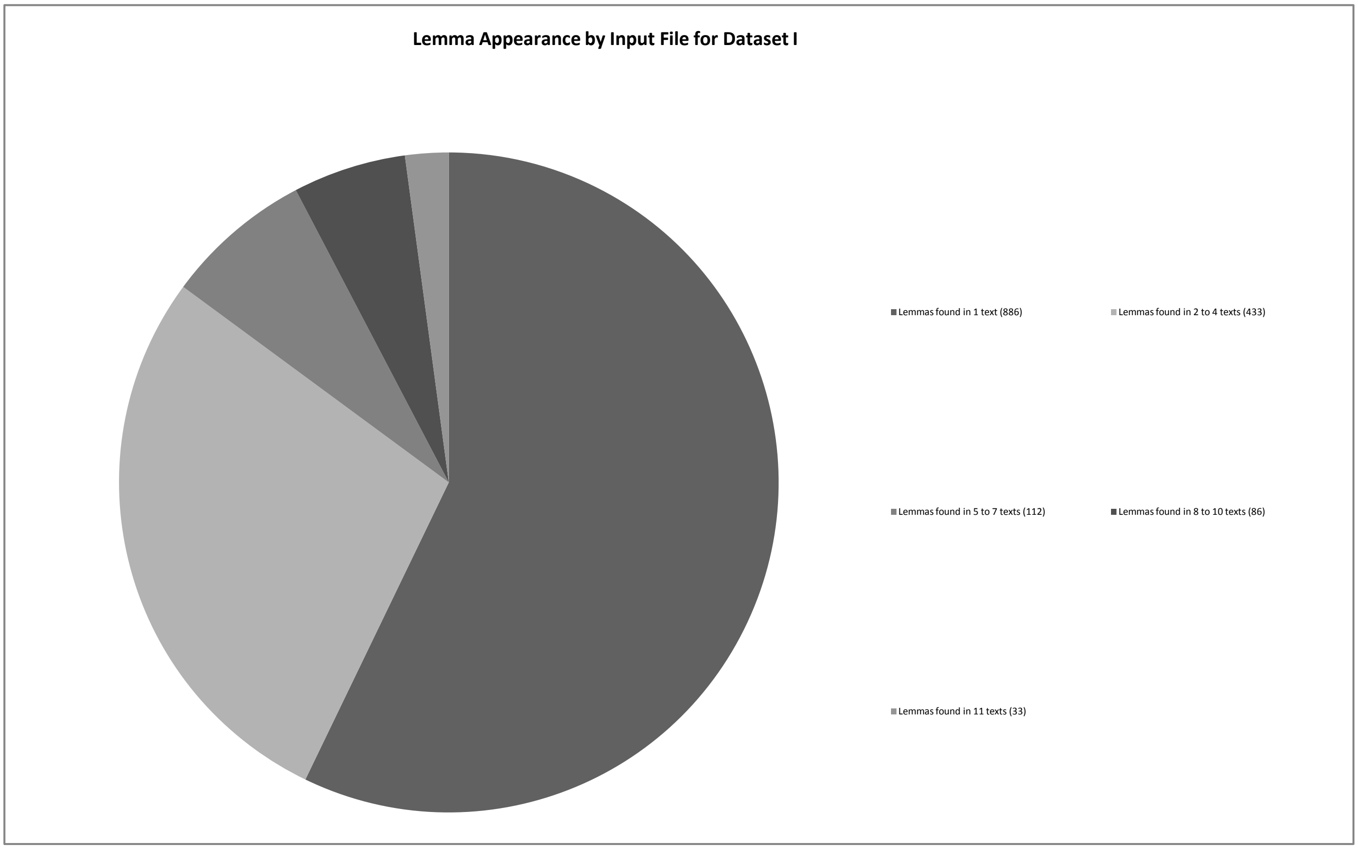
The average submission length for this group was 1197 words, (σ = 508). Carroll's type token ratio for this dataset is 9.55. The significance of this measure will be discussed with reference to the other two samples.

The primary goal of this analysis was to determine how many items overlap between texts. Figure VIII shows the distribution. A total of 33 lemmas were common to all eleven texts. These items are as follows*: ich, der, sein, haben, und, ein, mein, sich, in, mit, sehr, name, auf, zu, für, viel, Land, gehen, im, gut, um, nicht, sehen, auch, als, aber, können, über, Freund, essen, Kind, Jahr,*and *alt.* To repeat, the lemma "*ich*" represents all personal pronouns, the lemma "*mich*" stands for all reflexive pronouns like the lemma "*sich*" represents all reflexive pronouns. Similarly, "*sein*" represents all forms of the verb to be as "*mein*" represents all possessive adjectives.

These 33 lemmas made up 52.56% of the texts. While this number appears very large, one must remember that these 33 lemmas represent the derivatives of several different word groups. These items common to all eleven submissions were analyzed against the three vocabulary lists. This step was carried out to determine the source of the common vocabulary. The analysis revealed that all items were from the *Vorsprung* vocabulary list.

Considering that all students were assigned the same tasks, this number is rather low and is taken as proof that students in this dataset greatly varied their vocabulary.

Figure VIII: Lemma Appearance by Text for Dataset I



As is evident, most of these common 33 lemmas are function rather than content words. This signifies that the submissions overlapped mostly at the basic structure of each text rather than in the content, despite the fact that all students set out to complete the same exercises. Such a low number of items common among students further signifies one of two things: Firstly, one can argue that *Vorsprung* offers a large variety of vocabulary to students, allowing for creativity and uniqueness in their writing. Therefore, students were capable of answering all the required questions that were assigned but would still vary the vocabulary they use. Given that only 33 lemmas overlapped, one can further speculate that since such a large number of lemmas are left in the pool, the students were proficient in seeking out other sources of vocabulary such as dictionaries or thesauri.

My attention then turned to the 20th, 30th and 40th percentile of the items used. These items occur in two, three and four submissions. These 86 lemmas were compared to the three vocabulary lists and the analysis showed that 95.4% (82 lemmas) can be traced back to the *Vorsprung* vocabulary list, lending more support to the argument that *Vorsprung* provides learners with variable and frequent vocabulary. 3.5% (3 lemmas) and 1.1% (1 lemma) come from the Subtitles and *PDW* vocabulary list respectively.

A more detailed look at the lemmas in these percentiles allows for the rough categorization into the semantic fields of living arrangements, sports and hobbies. This list contains items students used to describe their current living arrangements such as the names of several rooms in a house, both the lemmas "*Katze*" [cat] and "*Hund*" [dog] – two animals prominently featured students' submissions as well as the names of some furniture items. This portion of the lemmas also included the names of two sports ("*Fußball*" and "*Eishockey*" [soccer and hockey]) and several lemmas relating to other sports. Lastly, several items from this list can be related to the theme of hobbies and extracurricular activities. The remaining lemmas in this category were either function words or items not pertaining to any noticeable category.

When the 50th, 60th and 70th percentiles were analyzed in more detail, a similar pattern emerged as in the preceding paragraph with reference to which vocabulary list these items can be traced back. Out of the 79 lemmas in this group, 93.6% (74 lemmas) were from the *Vorsprung* vocabulary list while 1.5% (1 lemma) and 2.5% (2 lemmas) come from the Subtitle and *PDW* list respectively. Only two semantic fields – family and occupation – are prominent in this section. The remaining lemmas are not necessarily related to any specific semantic fields.

Interestingly, the lemma "*heißen*" occurs in this percentile for the first time and is common to only seven out of the eleven submissions. This was particularly surprising since one of the first tasks of the term called for students to introduce themselves.

It is at this percentile interval that the first items not on any list appear. These items make up 2.5% (2 lemmas) and do not belong to just one semantic field. These lemmas are "*Sessel*" [armchair] and "*kämmen*" [to comb].

The remaining three percentile intervals were examined more closely in order to see if any patterns or semantic fields can be found. A total of 433 lemmas make up the last three percentiles. Again, the majority of these items came from the *Vorsprung* vocabulary list and made up 72% (312 lemmas). Roughly 8% (34 lemmas) were from the Subtitles list while 12% of these lemmas are from the *PDW* vocabulary list. The remaining 8% (34 lemmas) are not found in any vocabulary list.

A closer look at these items proved to be puzzling. The items "*Fakultät*" [department] and "*Aufsatz*" [essay] were not recognized by any list. An examination of the glossary of *Vorsprung* revealed that these items are indeed not an oversight but rather are not included in the textbook. One can conclude that these two items are not prominently featured in conversation in movies or stories newspapers either, but appear to be relevant to the cultural content of the Canadian university system.

The majority of the remaining items not found on any list could not be categorized in any one particular semantic field and were found to be compound words, like "*Schmerztablette*" [painkiller], "*Verkehrssystem*" [transportation system] and "*Berufserfahrung*" [work experience].

Lastly, the 886 words that occurred only once within these students' submissions were examined. Firstly, it is worth noting that this number is fairly high and it can therefore be argued that students at this level varied their vocabulary greatly when completing their submissions. The 886 items that occurred only once in the entire pool of the students' submissions were also compared to the three word lists to see to which word list they can be attributed. The analysis revealed that 32.7% (290 words) of these lemmas can be traced to the *Vorsprung* data, 6.3% (57 words) are found in the Subtitles vocabulary list, while 23% (206 words) of these lemmas are from the *PDW* vocabulary list. The remaining 37.5% (333 words) used within this group were not found in any of the word lists.

One conclusion that can be drawn from the findings generated from this sample is that despite being more advanced in the German studies the students still relied heavily on the vocabulary presented in their textbook. This is by no means a criticism of the student body, but rather clashes with the results I expected to see. However, this reliance on vocabulary from *Vorsprung* can also be viewed as positive as it illustrates that the *Vorsprung* textbook offers a large variety of vocabulary, given that a large proportion of items used only once stems from this source.

The next step in this analysis was the calculation of the average number of words used from each list in this dataset. The mean score from the *Vorsprung* vocabulary list was 1091.9 words, and 20.7 and 44.3 words from the Subtitles and WPD vocabulary list in that order. This is further proof that the students relied predominantly on words from the *Vorsprung* vocabulary list. The next analysis carried out on this dataset was a single factor ANOVA to determine if the number of items the learners used from the three vocabulary lists was statistically different. The number of words that students' received in all of the three vocabulary lists was included in this analysis. The analysis revealed that the difference in the number of words from each of the three lists was statistically significant as well (F (2, 30) = 63.16, p = 0.00.). This result further supports the argument that the students in this dataset did not rely on words from all three vocabulary lists evenly.

Next, a two sample t-test was carried out between the scores from the *Vorsprung* vocabulary list and the Subtitles vocabulary list and was not surprisingly, highly significant: t (10) = 2.2, p = 0.01. This test further supports the finding that the *Vorsprung* vocabulary list is the dominant source of vocabulary for students.

The same pattern was found once a two sample t-test was carried out on the students' scores from the Subtitles and *PDW* vocabulary list: t (10) = 2.17, p = 0.04. The results from these two t-tests show that the differences in the number of words that were traced back to the three vocabulary lists differed significantly from each other, which further allows for the conclusion that the lemmas found in these three lists were used in differing amounts by the students in this dataset.

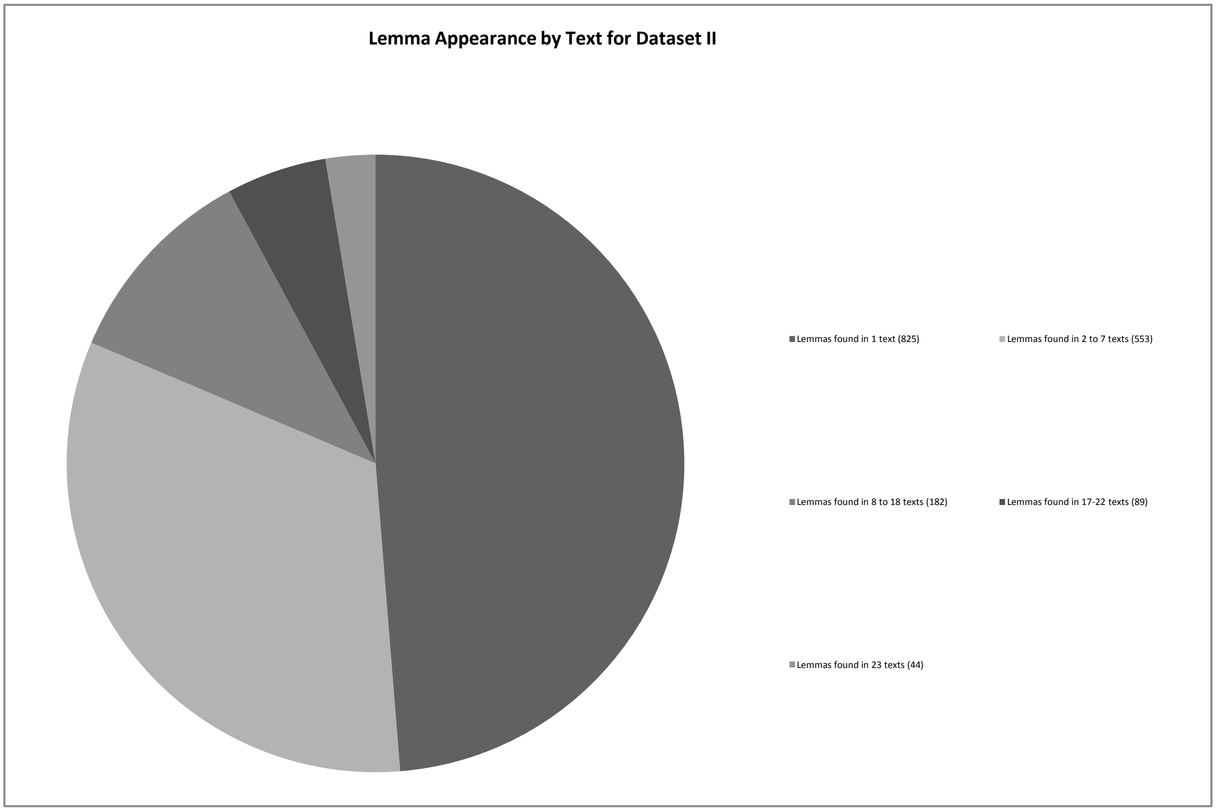
I argue that the findings illustrate that although they rely heavily on the vocabulary from the *Vorsprung* list, students still varied their vocabulary use while completing their submissions. Although intriguing on their own as well, the findings drawn from with this dataset become notably more interesting once considered in comparison with the next two datasets. The findings from dataset II will be discussed next.

5.2. Dataset II: Analysis of the Submissions for German 102

The submissions of all 23 students from German 102 were analyzed all together first. The average text length for this group was 1195 lemmas (σ = 225) and Carroll's type token ratio was 7.23. The average text length does not differ greatly from the average text length in dataset I. A more noticeable difference between this and the aforementioned sample can be seen at the scores obtained for Carroll's type token ratio, with the ratio for this dataset being lower that the ratio in dataset I. I take this to be a sign of lexical variation influenced by group membership, in this case meaning that students in the more advanced course varied their vocabulary more than students in German 102.

This analysis of comparing all students' submissions to each other also revealed what the common vocabulary between all submissions is. Word appearance by input file for this sample is illustrated in Figure V: Word Appearance by Input File. The lemma occurrences were again presented in percentiles. The largest proportion of this chart is taken up by words that occurred only once within this sample. This is also taken to illustrate a large variability of vocabulary use within this group.

Figure IX: Lemma Appearance by Text for Dataset II



For this group, a total of 44 common lemmas occur in every submission. These lemmas are as follows: *aber, auch, aus, Bank, besuchen, Bier, der, dies, dürfen, essen, ein, fahren, Familie, für, gehen, gut, haben, Haus, ich, im, in, kommen, können, Land, mein, mit, mögen, müssen, nach, nicht, oder, sehr, sein, sich, sollen, Stadt, studieren, und, viel, von,* *wann, wenn, Winter,* and *zu.* An analysis using *Range* revealed that all of these items come from the *Vorsprung* textbook. An analysis with *Frequency* revealed that these lemmas take up 58.17% of all of the submissions from this dataset.

Again, as all students were assigned the same tasks, one can argue that 44 is a low number for overlapping lemmas between all submissions. This can also be taken as another sign of variability in vocabulary use. It is also important to point out that the variety of word classes within these lemmas such as nouns, verbs and prepositions to name a few, are indicators of variability in both vocabulary type and presumably sentence length.

Next, my attention turned to the lower quarter of the lemmas used within this group. These lemmas appeared in 2 and 7 input files. There were 553 lemmas in this group. *Range* revealed that 61.8% (342 lemmas) come from the *Vorsprung* vocabulary list. The remaining 9.2% (51 lemmas), 12.3% (68 lemmas) and 16.6% (92 lemmas) were from the Subtitle vocabulary list, *PDW,* and not from any list, respectively. The semantic fields covered include but are not limited to weather, living arrangements and hobbies.

The next analysis was carried out on the items that appeared between 17 and 22 students' files. This categorisation roughly represents the 75th percentile of vocabulary use within this sample. This section consists of 89 lemmas in total. *Range* indicated that 97.7% (87 lemmas) of this section come from the *Vorsprung* vocabulary list while the remaining 2.2% (2 lemmas) come from the *PDW* vocabulary list. No items at this section come from the Subtitles vocabulary list. A closer analysis of the lemmas in this section allows for the categorisation into a few semantic fields such as family, living arrangements, and hobbies.

All three of these themes can be tied back to the course syllabus as students were asked to write about these topics.

One of the two lemmas from the *PDW* vocabulary list was "*Alptraum*" [nightmare]. The same concept of nightmare is present in the *Vorsprung* vocabulary list but is spelled as "*Albtraum*" and was therefore not recognised as belonging to any vocabulary list. Both spellings of nightmare are acceptable in German but the fact that the word occurs in more than one list is a mistake I made. Because the concept nightmare is in fact present on the *Vorsprung* vocabulary list, this leaves only one lemma – "*Tradition*" [tradition] from the *PDW* vocabulary list.

Lastly, a closer examination was also carried out on the rarely used items within this sample, namely the 825 lemmas that occurred only once within this group. These lemmas were examined in more detail as they – if considered as a group – take up slightly less than an average submission by a student.

The majority of these items, namely 45.6% (376 lemmas), were not in any vocabulary list. A large proportion of these items are compound words such as "*Abenteuersportarten*" [extreme sports], "*Busenfreundin/Seelenfreundin*" [soul mate], "*Geburtstagsgeschenk*" [birthday gift], "*Geschwindigkeitsbegrenzung*" [speed limit] or "*Geschwindigkeitsbeschränkung*" [speed restrictions]. All of these lemmas are accurate words from the German language and some of their stems when considered in isolation can be traced back to the *Vorsprung* vocabulary list. Once combined with other stems, these items fail to show up in said list. The presence of these lemmas within this sample cannot therefore be attributed to knowledge acquired through the textbook. One must speculate that these lemmas were either obtained through the use of a dictionary or similar learning resources or through help obtained from a more knowledgeable speaker of the German language.

Two prominent semantic fields were noted amongst the items that were not recognised by any word list. These are living arrangements and animals. Items like "*Abstellraum*" [storage room], "*Apartmenthaus*"[apartment house], "*Autogarage*" [car garage] or "*Bettlaken*" [bed sheets], just to name a few, were clearly deemed important enough by a student to use and better describe his or her living arrangements, but were not amongst the basic vocabulary needed to do so.

Furthermore, several animals were present such as "*Bär*" [bear], "*Kaninchen*" [rabbit] and "*Ente*" [duck], and could possibly be traced back to submissions about Canada as one of the exercises called for students to write about what visitors can expect to see and do when they travel to Canada.

Next, the same statistical analyses as were carried out on dataset I were applied to this sample as well. The mean word use from the three vocabulary lists was 1109 words, 14.5 words and 28.5 words from the *Vorsprung,* Subtitles and *PDW* vocabulary lists respectively. The ANOVA calculated on the scores of the students of this dataset also revealed that the number of words traced back to the three vocabulary lists differed from each other: (F (2, 63) = 617.51, p = 0.00.). As in the previous sample, the t-tests carried out between the *Vorsprung* vocabulary list and the Subtitles list as well as the comparison between the scores on the Subtitles vocabulary list and *PDW* list were significant: t (21) = 2.01, p = 0.00; t (34) = 2.03, p = 0.00. This is taken as further proof that students did not rely on all three vocabulary lists evenly, but rather that the number of lemmas from each vocabulary lists differs statistically. As in the previous dataset, students relied mostly on the vocabulary presented in the *Vorsprung* textbook but also used words that can be classified as belonging to the other two lists.

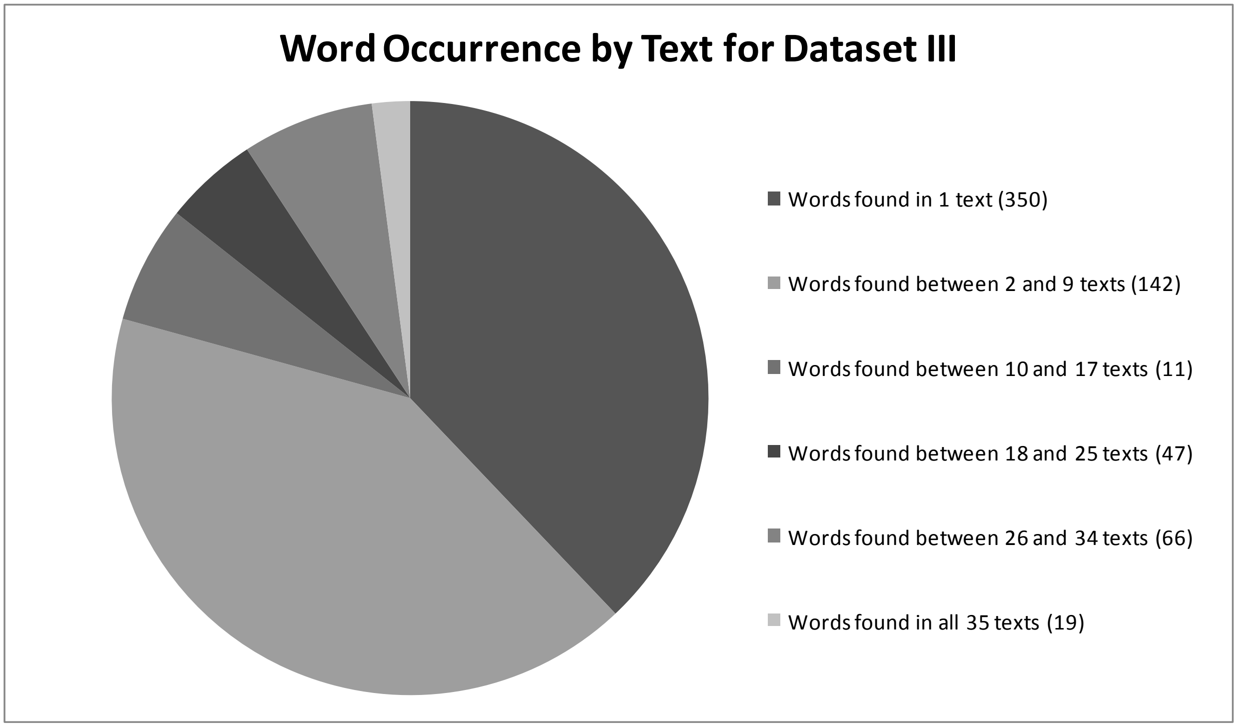
Having completed an analysis of the first two datasets, the my focus will now shift to the third and final dataset. Following this, comparisons between all three datasets will be made.

5.3. Data Set III: Analysis of the Submissions for German 101

The submissions of all 35 German 101 students were analyzed with reference to each other. This was done to extract the common vocabulary between the works of all students in this sample. The average submission length for this group was 645 words. The type token ratio for this dataset is 4.1. Already through these descriptive statistics one can see a difference between this dataset and the earlier two datasets. The average text length of this dataset is noticeably shorter the average text length in the other datasets. The score on Carroll's type token ratio the students in this course obtained is also markedly lower than the other two samples.

Again, an analysis was carried out to determine the pattern of word use within this sample and see which words were common to all submissions. Because of the large sample size for this data set, the submissions and corresponding word occurrences were placed into groups. The word occurrence by text can be seen in Figure X: Word occurrence by text for Dataset III.

Figure X: Word Occurrence by Text for Dataset III



Only 19 lemmas were common to all 35 submissions. These items are as follows: *alt, aus, der, ein, gehen, gut, haben, hier, ich, in, kein, kommen, mein, nicht, sehr, sein, sprechen, und*, and *wie*. It is also worth noting that no nouns are featured in this list of overlapping lemmas despite the fact that articles are represented. Not surprisingly, all 19 items are also found on the *Vorsprung* vocabulary list. An analysis with *Frequency* revealed that this group of lemmas covers 49% of the submissions in this dataset.

Attention must be drawn to two points associated with the number of overlapping lemmas in this sample. Firstly, this low number deviates from the pattern observed in the previous two datasets. To repeat, in the datasets I and II a total of 33 and 44 lemmas were common to all submissions respectively. The assumption I held prior to the analysis is that the number of overlapping lemmas between all texts in this sample would be larger than the numbers observed in datasets I and II. Given that the learners in this sample have lower familiarity with the German language and its vocabulary, I speculated that this sample would show a larger number of overlapping lemmas. But one must keep in mind that the average submission for this sample was significantly shorter and therefore, one can expect the overlapping number of lemmas to be smaller as well. Furthermore, although the number of overlapping lemmas in this sample is smaller than the number of overlapping lemmas observed in the above discussed datasets, these 19 lemmas still take up roughly the same amount of text space as the overlapping lemmas in the previous two samples. To repeat, these percentages were 52% and 58% in dataset I and II respectively, while in this sample the 19 lemmas take up 49% of the texts.

The first analysis was carried out on the rarer items occurring in between 2 and 9 texts. Within this group, 69.9% (259 lemmas) come from the *Vorsprung* vocabulary list. The remaining lemmas taking up 6.7% (25 lemmas) and 10.7% (40 lemmas) can be found in the Subtitles and *PDW* lists respectively.

The last proportion making up 12.9% (48 lemmas) was not featured in any vocabulary list. Once examined more closely, it was evident that a large number of these words relate to the semantic field of food, which can explain why they are not featured in any of the vocabulary list. After all, "*Ahornsirup*" [maple syrup] is presumably not a frequently used German word. Another frequently observed construction is the use of "favorite" as in

"*Lieblingsonkel/tante/nichte*" [favourite uncle/aunt/niece], which is another construction not covered in the vocabulary lists used for comparison.

An assessment of the items occurring between 10 and 17 texts was then carried out. This group consisted of 59 words. Out of these 59 words, 96.6% (57 lemmas) could be traced back to the *Vorsprung* vocabulary list. Color names were frequently in this group of lemmas, as are a few names for family members and foods. The remaining lemmas cannot be placed into specific semantic fields.

Only one item in this group of lemmas came from the Subtitle list, and only one item was not found in any list. These items were "*wach*" [awake] and "*Hockey*" [hockey] respectively. Each of these two items took up 1.7%. The presence of the term "Hockey" in the students' submission is unsurprising, given the popularity the sport enjoys in Canada. Its absence from the word lists used for comparison is also common sense, given the lack of popularity the sport enjoys in Germany.

The next section analysed was lemmas that were used in between 26 and 34 texts. The analysis revealed that all 66 lemmas from this group came from the *Vorsprung* vocabulary list. Semantic fields covered in this group of lemmas are education and school supplies, family members and living arrangements.

The final analysis carried out on this dataset was on the items that occur only once. This group consisted of 350 lemmas. 39.4% of these lemmas (138 lemmas) were from the *Vorsprung* vocabulary list. The remaining 11.4% (40 lemmas), 20.6 % (72 lemmas) and 25.7% (100 lemmas) were from the Subtitles vocabulary list, the *PDW* vocabulary list and not featured on any vocabulary list in that order. One noticeable feature amongst the lemmas from the *Vorsprung* and the Subtitles vocabulary list is that many of the items are adjectives.

A large number of the words not found in any vocabulary list related to food. Again, many of the items not found on any vocabulary list were compounds such as "*Orientierungspunkt*" [point of orientation] or "*Sonnenterasse*" [sun terrace].

The mean word use from the three vocabulary lists for this dataset was 623 words, 4.16 and 7.7 words from the *Vorsprung*, Subtitles and *PDW* lists correspondingly. As with the previous two datasets, a single factor ANOVA was carried out to determine if further analysis should be carried out with the scores of this dataset. This analysis revealed that as in the previous two datasets, the difference in lemma use from the three vocabulary lists was significant: F (2, 105) = 734.65, p = 0.01. The t-test between the *Vorsprung* vocabulary list and the Subtitles vocabulary list was also significant: t (35) = 2.03, p = 0.00. Similarly, the t-test between the Subtitles vocabulary list and the *PDW* list was also significant: t (59) = 2.00, p = 0.00.

The next section will present a discussion evolving from a comparison between these three datasets.

5.4. Comparison between the three Datasets

A few descriptive statistics will be discussed first in the comparison between the three datasets. I felt that comparing the three datasets is of interest because it allows for a discussion on how the three datasets differ.

Ellis and Yuan (2004) argued that learners of lower proficiency devote significantly more time to lexical aspects of their writing, often at the expense of text length. If considering average text submissions, this finding is true for the datasets in study. To repeat, for dataset I, II and III the mean text submissions were 1197, 1195 and 645 lemmas in that order. There is considerable change in text length between students in the introductory course and the upper two courses. This difference disappears at the higher level courses within these samples as evidenced by the almost identical mean submission lengths between dataset I and II.

The three datasets also differed in their score on Carroll's type token ratio. To reiterate, the scores were 9.55, 7.23 and 4.3 for dataset I, II and III in that order. These scores demonstrate that lexical variety is influenced by course membership and increases as students progress in their language learning process. Put differently, students who are just beginning to learn the language were more likely than students in more advanced levels of language learning to repeat the same words in their submissions. One can also rephrase this findings by stating that students with more advanced linguistic skills were more likely to use a larger variety of words when completing their assigned work.

All three datasets relied heavily on the vocabulary from the *Vorsprung* vocabulary list. They differed in their use of words found in the other two lists. The data in this study show that as students progressed in proficiency as marked by their more advanced language courses, the number of lemmas from the *PDW* vocabulary list increased. To reiterate, the average submission of a German 101 student featured 7.7 words from the *PDW* vocabulary list, while this number was 28.5 words and 44.3 words for the average German 102 and 201 student respectively. Given that in this research study, the *PDW* vocabulary list acted as a list of less frequent vocabulary, the data illustrate that students with more advanced knowledge of the German language were more likely to use less frequent words. This in turn is similar to the findings of Laufer and Nation's original study.

Within the three datasets discussed above, the number of lemmas that occur only once also increased as students progressed in courses. As a reminder, datasets I, II and III contained 886, 825 and 361 words that occurred once in that order. Put differently, the students were more likely to select rare words to include in their writing if they were in more advanced stages of language learning.

Next, the analysis concentrated on determining what the common vocabulary between the three datasets is. The common lemmas that occurred within all submissions of one sample were compared to each other. To clarify, the 33 lemmas common to all German 201 submissions, the 44 lemmas common to all German 102 submissions and the 19 lemmas common to all German 101 students were analyzed by *Range.*

The analysis revealed that 12 lemmas were occurred in every single submission analyzed in all three datasets. These lemmas are as follows: *der, ein, gehen, gut, haben, ich, in, mein, nicht, sehr, sein* and *und*. As is evident, no nouns are found in this group. Furthermore, these lemmas cannot be placed into one specific semantic field either. These lemmas cover a large band of grammatical categories such as definite and indefinite articles, possessive pronouns and a few key verbs. Put differently, no specific content can be communicated relying solely on the above printed lemmas. But in order to write a coherent text – regardless of proficiency or topic – many of these items would have to be used.

The next analysis focused on the rare words in each dataset. More specifically, I was interested in knowing how many of the items that occur once in every dataset – the rare words – occur in the other two datasets as well. In datasets I, II and III, 886, 825, 361 lemmas occur only once respectively.

However, between these three sets, only 7 rare words overlap. These words are: *Bein, darüber, einmal, empfehlen, ideal, innerhalb,* and *Nachmittag*. Again, no one semantic field is covered by these seven lemmas. All but one of these lemmas belong to the *Vorsprung* vocabulary list while "*ideal*" [ideal] was traced back to the *PDW* vocabulary list.

All three datasets had a large proportion of words that I could not account for through any vocabulary list. In Laufer and Nation's original work relying on this framework, these items are considered rare words and were taken to signal an increasing vocabulary size. Although some of these items not accounted for in this study could indeed be infrequent words in the German language, I hesitate to label *all* of the lemmas not accounted for by any vocabulary list as rare because: Several of the items not found in any vocabulary list are compound words. In many cases, roots of these compound words can be traced back to one of the three vocabulary lists. This in turn can be interpreted in one of two ways: The learners could be demonstrating greater familiarity with the German language by combining words in their writing. After all, given that the English language does not allow for such constructions, this is a novel feature for the learners and the presence of such compounds is a sign of language learning. Thus although a large proportion of lemmas was not accounted for in this study, they cannot be taken to be very infrequent words as many of these lemmas were simply not found in the vocabulary lists because of their compound nature. Regardless of what the reason for the high proportion of compounds is, the learners are demonstrating great variability in the vocabulary they use.

Having introduced some qualitative and quantitative differences between the datasets, I will now discuss the implications of these findings as well as offering conclusions based on the data and suggestions for further research.

Chapter 6

Conclusions and Suggestions for Future Research

6.1. Conclusions

One goal of this work was to evaluate if differences exist between texts by students of the three proficiency levels of proficiency as measured in the way learners use vocabulary in their written submissions. As the above presented discussion showed, differences were noted between the three datasets. To reiterate briefly, the first two datasets featured longer texts on average than the third dataset. Similarly, Carroll's type token ratio scores increased as students progressed from the introductory course to the intermediate course. The overlapping lemmas in German 201 and 102 featured more lemmas than in German 101. In all three datasets discussed, a large proportion of the lemmas could be traced back to the *Vorsprung* vocabulary list, indicating that the learners used the words they acquired in their language classes most frequently. Furthermore, in all three datasets, there was significant difference in the number of words that came from the three vocabulary lists used for comparison. These findings were validated and therefore the arguments strengthened through the use of qualitative tools. I believe to have demonstrated that as students progress in their language courses, their vocabulary use and associated frequency change to reflect their increasing familiarity of the German knowledge. One of the two goals for this work has been met.

The other goal of this work was to evaluate the usefulness of lexical frequency profiling within this population and its adaptation to the German language. Several factors must be taken into account when evaluating this tool. Firstly, both *Range* and *Frequency* were very easy to employ. Both software programs proved to be very informative tools in evaluating the students' texts, in particular because of the type of information these pieces of software provide to the researcher. Another benefit to these software programs is the speed at which results can be obtained. A few clicks offered a wealth of information on any text analyzed. In addition, the output and results were very easy to interpret and share. The results gained show signs of different vocabulary use and creativity in the way vocabulary was utilized by students. Without a doubt, these two tools can be of great service to anyone wishing to examine lexical use.

However, in order to get to the benefits of this study and adapting it for use with the German language, a disproportionate amount of time was invested in preparing all the materials to align them with *Range* and *Frequency*. This is by no means a fault of the software programs, but rather a characteristic of the German language. Although I take great pride in the outcome of this study, the difficulty in executing it to obtain the maximum results cannot be overlooked. To repeat, all submissions had to be stripped of several word groups before the text could be lemmatized. These lemmatized outputs had to be further cleaned to be analyzed for lexical frequency. In addition, all vocabulary lists had to be standardized and quite specifically customized for this study. In the end, a significantly longer amount of time was needed to carry out all these early steps than it took to analyze the data.

Furthermore, despite the fact that several measures were taken to adopt Laufer and Nation's (1995) procedure and approach towards lexical frequency profiling, this work fell short in one dimension. As evidenced in the above presented results section, I was unable to account for a large number of lemmas that were utilized by students. Laufer and Nation's results feature a significantly smaller number of words that were unaccounted for. Additionally, the words Laufer and Nation could not account for could all be neatly classified as rare words, thus demonstrating lexical growth by the learners. This thesis project could not enjoy the same benefit as the original study as I could not argue that all words unaccounted for are infrequent words.

One factor associated with this is the frequency of compound words in the German language. A large number of the words I was not able to account for were such compound words, all of which were grammatically correct, but would be infrequent outside of the classroom setting. One such example used by a student is the item "*Universitätsbibliothek*" [university library]. Both "*Universität*" [university] and *Bibliothek*" [library] can be found in the *Vorsprung* list, but their combination cannot. Language specific characteristics such as these must be considered when implementing this framework.

In retrospect, such compound words could have been split before the analysis with *Range* was carried out. Doing so would have allowed me to account for more items through the vocabulary lists. But by splitting these compound words, I would be changing the students' submissions even further from their original context. Additionally, splitting up a term like "*Universitätsbibliothek*" into "*Universität*" and "*Bibliothek*" also takes away from the fact that the learner had enough experience with the German language to combine the two stems in a grammatically correct way.

Another aspect to keep in mind when implementing the language frequency profile is the order in which the vocabulary lists are arranged. Laufer and Nation also used three lists, but their lists reflected frequency of words in the English language. Thus it was logical for the authors to place the list containing the 1000 most frequent word families of the English language in first position and the list containing the second 1000 most frequent word families of the English language in second position.

Since this project relied on generated lists or modified versions of already existing word frequency lists, the order in which these lists were organized was based on assumptions I had before analyzing the data. Assuming that learners would use more words from their textbook, the *Vorsprung* vocabulary list was placed in first position. Furthermore, I assumed that the least number of words used would come from the *PDW* vocabulary list because of the high standard of German used in the corpus on which this list was based. For this reason, the *PDW* vocabulary list was placed in third position, leaving the Subtitles vocabulary list in second.

A glance at the data revealed that for all three datasets, the number of lemmas from the *PDW* vocabulary list exceeds the number of lemmas from the Subtitles list. If employing the lexical frequency profile again with the three vocabulary lists used in the present work, it would be of more benefit to reverse the order of the Subtitles and *PDW* vocabulary lists, as the data suggest that students were more likely to draw on words from the *PDW* list than the Subtitles list. Similarly, one could place the *Vorsprung* vocabulary list in third place and rank the other two lists in first and second place. Doing so would approximate Laufer and

Nation’s (1995) design even further. These are unfortunately lessons and ideas that can only be learned in retrospect and once all the data are analyzed, but it are worth keeping in mind for future implementations of the lexical frequency profile.

To conclude, despite the challenges faced in accommodating characteristics of the analyzed texts with the lexical frequency profile, this framework was certainly a worthy and reliable tool for this analysis and can be utilized successfully in future studies that examine similar factors of vocabulary use.

6.2. Critique of the present work

Collentine (2004) writes that any comprehensive approach to second language acquisition needs to consider both internal and external factors that interact with and influence acquisition of a foreign language. While this is sound logic and certainly good advice, this analysis is limited to only the output resulting from said internal and external factors. As mentioned in the introduction and evidenced throughout the body, this work focuses solely on the production of written text by learners of German without consideration for metacognitive aspects associated with learning. No attempt was made to acquire further information about the students’ learning processes such as the amount of time they invested studying the course materials, their learning strategies, the knowledge of the German language students had before enrolling in these courses or the grades the students received. Therefore, no conclusions past the group level can be made for the data presented above. The goal was not to conduct a longitudinal study across years of language learning but rather to examine the vocabulary production of learners during one term and to test if an approach relying on lexical frequency profiling can be used to better study vocabulary use within these three populations.

The present analysis can be used as a starting point for further exploration of vocabulary learning and can certainly be integrated with and supported by the inclusion of metacognitive factors. Furthermore, learners' metalinguistic knowledge and their linguistic proficiency have been found to correlate positively and significantly (Roehr, 2008). There is evidence for an overall association between higher levels of a learner's awareness of metalinguistic knowledge and their performance in the second language. While metalinguistic awareness does not guarantee successful performance in the target language, instructors and researchers would be well advised to foster this trait as it is associated with better acquisition. This is but one suggestion as to how this framework can be of benefit for learners of a foreign language. More suggestions will be discussed in the following section.

6.3. Suggestions for Future Research

This line of work can be continued at different proficiency levels and could be used to examine the differences in vocabulary use that exist between native speakers and highly proficient non-native speakers. However, for a project such as that one, different vocabulary lists would have to be used for comparison.

An approach such as the lexical frequency profiling can easily be added as a part of student modeling into learning software. Doing so could improve the quality of feedback learners receive while interacting with the software as it could tell a learner what kind of grade to expect before they submit something electronically.

In addition, researchers and instructors interested in further understanding and fostering vocabulary development would be well advised to examine the size and organization of learners’ vocabulary networks rather than attributes of individual words or vocabulary size in exclusion. Bygate (1999) writes that the completion of any one language task results in learning. Learning is not limited only to the target language but extends to learning how to process the language at hand. The author adds that language use and language processing cannot exist separately. As such, an examination should be carried out focusing on what kind of learning tasks foster the most vocabulary acquisition. Kim (2008) argues that tasks calling for large degree of generative processes meet these criteria as they call for a deeper level of processing. Tasks like these were examined in this work, but the learners in this sample are all still in relatively early stages of language learning. A similar examination could be carried out on more proficient learners of German.

To better understand vocabulary use in a writer’s foreign language, researchers would also be well advised to evaluate it against a baseline performance in the writer’s native language. Doing so would reduce effects of linguistic talent in the native tongue when evaluating the skills in a foreign language.

In addition to the aforementioned suggestions, any approach to the examination of vocabulary use by learners would also benefit from a more multidisciplinary approach. A wealth of findings from psychology coupled with established findings from linguistics could be used to better understand vocabulary learning, which would in turn enhance the way learners make use of their lexical skills.

Bibliography:

Alderson, C. J. (2007). Judging the frequency of English words. *Applied Linguistics, 28(3),* 383-409.

Auer, E. T. & Bernstein, L.E. (2008). Estimating when and how words are acquired: A natural experiment on the development of the mental lexicon. *Journal of Speech, Language and Hearing Research, 51,* 750-758.

Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing, 17(1)* 1-42.

Basanta, C. P. (2004). Pedagogic aspects of the design and content of an online course for the development of lexical competence: ADELEX. *ReCALL 16(1),* 20–40.

Biemann, C., Bordag, S., Heyer, G., Quasthoff, U. & Wolff, C. (2004). Language-independent methods for compiling monolingual lexical data. A. Gelbukh (Ed.): CICLing 2004, LNCS 2945, pp. 217–228, Springer-Verlag Berlin Heidelberg, 2004.

Brent, W. (2006). Lexical network structures and L2 vocabulary acquisition: The role of L1 lexical/conceptual knowledge. *Applied Linguistics, 27(4),* 741-747.

Bygate, M. (1999). Task as context for the framing, reframing and unframing of language. System, 27*,* 33-48.

Collentine, J. (2004). The effects of learning contexts on morphosyntactic and lexical development. *Studies in Second Language Acquisition, 26,* 227-248.

Corrigan, R. (2007). Dimensions of deep vocabulary knowledge used in inferring the meaning of words in context. *Applied Linguistics, 28(2),* 211–240.

Crais, E. R. (1990). Word knowledge to work knowledge. *Topics in Language Disorders, 10(3),* 45-62.

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

Daller, H., Van Hout, R., & Treffers-Daller, J. (2003). Lexical richness in the spontaneous speech of bilinguals. *Applied Linguistics 24(2),*197-222.

Davidson, D. J., Inderfey, P. Gullberg, M. (2008). Words that second language learners are likely to hear, read and use. *Bilingualism: Language and Cognition, 11(1),* 133-146.

Dewaele, J.M. (2008). Dynamic emotion concepts of L2 learners and L2 users: A second language acquisition perspective. *Bilingualism: Language and Cognition, 11(2),* 173-175.

East, M. (2004). Calculating the lexical frequency profile of written German texts. *Australian Review of Applied Linguistics, 27(1),* 30-43.

East, M. (2006). The impact of bilingual dictionaries on lexical sophistication and lexical accuracy in tests of L2 writing proficiency: A quantitative analysis. *Assessing Writing, 11,* 179-197.

Elman, J.L. (2004). An alternative view of the mental lexicon. *Trends in Cognitive Sciences, 87(7),* 301-306.

Ellis, R. & Yuan, F. (2004). The effects of planning on fluency, complexity and accuracy in second language narrative writing. *Studies in Second Language Acquisition, 26,* 59-84.

Faulstich, L. C., Quasthoff, U., Schmidt, F. & Wolff, C. (2002). Concept extractor - Ein flexibler und domänenspezifischer Web Service zur Beschlagwortung von Texten. In:

Hammwöhner, Rainer; Wolff, Christian; Womser-Hacker, Christa (Hg.): Information und Mobilität, Optimierung und Vermeidung von Mobilität durch Information. Proceedings des 8. Internationalen Symposiums für Informationswissenschaft (ISI 2002), Regensburg, 8. – 11. Oktober 2002. Konstanz: UVK Verlagsgesellschaft mbH, 2002. S. 165 – 180.

Fitzpatrick, T. (2007). Word association patterns: Unpacking the assumptions. *International Journal of Applied Linguistics, 17(3),* 319-332.

Heatley, A., Nation, P. & Coxhead, A. (2002). RANGE and FREQUENCY Programs. Retrieved August 1, 2008, from http://www.victoria.ac.nz/lals/staff/paul-nation/nation.aspx

Hover, D. L. (2003). Another perspective on vocabulary richness. *Computers and Humanities, 37,* 151-178.

Huibregtse, I., Admiraal, W., & Meara, P. (2002). Scores on a yes-no vocabulary test: Correction for guessing and response style. *Language Testing, 19(3),* 227-245.

Ijaz, I.H. (1986). Linguistic and cognitive determinants of lexical acquisition in a second language. *Language Learning, 36(4),* 401-451.

Kim, Y. (2008). The role of task-induced involvement and learner proficiency in L2 vocabulary acquisition. *Language Learning, 58(2),* 285-325.

Larsen-Freeman, D. (2006). The emergence of complexity, fluency and accuracy in the oral and written production of five Chinese learners of English. *Applied Linguistics, 27(4),* 590-619.

Laufer, B. (1991). The development of L2 lexis in the expression of the advanced learner. *The Modern Language Journal, 75(4),* 440-448.

Laufer, B. & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics, 16(3),* 307-322.

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

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

Laufer, B. (2005). Lexical frequency profiles: From Monte Carlo to the real world. A response to Meara (2005). *Applied Linguistics, 26(4)*, 582-588.

Leach, L. & Samuel, A. G. (2007). Lexical configuration and lexical engagement: When adults learn new words. *Cognitive Psychology, 55*, 306-353.

Lezius, W., Rapp, R., & Wettler, M. (1998). A freely available morphological analyzer, disambiguator, and context sensitive lemmatizer for German. *Proceedings of the COLING-ACL 1998*, 743-747.

Lovik, T. A.L, Guy, J.D. & Chavez, M. (2007). *Vorsprung: A Communicative Introduction to German Language and Culture.* Houghton Mifflin Company, Boston: New York.

Meara, P. (2004). Modelling vocabulary loss. *Applied Linguistics, 25(2),* 137-155.

Meara, P. (2005). Lexical frequency profiles: A Monte Carlo analysis. *Applied Linguistics, 26(1)*, 32-47.

Nation, P. (2001). *Learning Vocabulary in another Language*. New York: Cambridge University Press.

Nation, P. (2005). Range. [Computer software]. Wellington, New Zeland: Victoria University of Wellington. Retrieved September 21, 2008. Available from: http://www.victoria.ac.nz/lals/staff/paul-nation/nation.aspx

Nation, P. (2005). Frequency. [Computer software]. Wellington, New Zeland: Victoria University of Wellington. Retrieved September 21, 2008. Available from: http://www.victoria.ac.nz/lals/staff/paul-nation/nation.aspx

Read, J. (2000). *Assessing vocabulary*. New York: Cambridge University Press.

Quasthoff, U. & Wolff, C. (1999). Korpuslinguistik und große einsprachige Wörterbücher. *Linguistik Online, 3.*

Quasthoff, U., Richter, M. & Biemann, C. (2006). Corpus portal for search in monolingual corpora. Proceedings of the fifth international conference on Language Resources and Evaluation, LREC 2006, pages 1799– 1802, 2006.

Read, J. & Chapelle, C. A. (2001). A framework for second language vocabulary assessment. *Language Testing, 18(1),* 1-32.

Roehr, K. (2008). Linguistic and metalinguistic categories in second language learning. *Cognitive Linguistics, 19(1),* 67-106.

Schoonen R. & Verhallen, M. (2003). The assessment of deep word knowledge in young first and second language learners. *Language Testing, 25(2),* 211-236.

Schulze, M., Liebscher, G. & Su, M. Z. C. (2007). *Geroline: S*tudent perception and attainment in an online German language course. *German as a Foreign Language, 1,* 1-25.

Schulze, M., Wood, P., & Pokorny, B. (2009). Measuring Balanced Complexity in Online Writing. Manuscript submitted for publication.

Tavakoli, P. & Foster, P. (2008). Task design and second language performance: The effects of narrative type on learner output. *Language Learning, 58(2),* 439-473.

Webb, S. (2008). Receptive and productive vocabulary sizes of L2 learners. *Studies in Second Language Acquisition, 30,* 79-95.

Wiktionary: Top 2000 German words from subtitles. (2009). Retrieved January 12, 2009, from: Wiktionary: http://en.wiktionary.org/wiki/Wiktionary:Frequency_lists#German

Wolter, B. (2001). Comparing the L1 and L2 mental lexicon. *Studies in Second Language Acquisition, 23(1),* 41-69.

Zareva, A., Schwanenflugel, P. & Nikolova, Y. (2005). Relationship between lexical competence and language proficiency. *Studies in Second Language Acquisition, 27,* 567-595.