QuickAssist
Extensive Reading for Learners of German Using CALL Technologies

by

Peter Wood

A thesis

presented to the University of Waterloo

in fulfilment of the

thesis requirement for the degree of

Doctor of Philosophy

in

German

Waterloo, Ontario, Canada, 2010

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 the examiners.

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

Abstract

The focus of this dissertation is the development and testing of a CALL tool which assists learners of German with the extensive reading of German texts of their choice. The application provides functionality that enables learners to acquire new vocabulary, analyse the meaning of complex word forms and to study a word's semantic and syntactic features with the help of corpora and online resources.

It is also designed to enable instructors to create meaningful exercises to be used in classroom activities focusing on vocabulary acquisition and word formation rules.

The detailed description of the software development and implementation is preceded by a review of the relevant literature in the areas of German morphology and word formation, second language acquisition and vocabulary acquisition in particular, studies on the benefits of extensive reading, the role of motivation in second language learning, CALL, and natural language processing technologies.

The user study presented at the end of this dissertation shows how a first test group of learners was able to use the application for individual reading projects and presents the results of an evaluation of the sortware conducted by three German instructors assessing the affordances of the applications for students and potential applications for language instructors.

Acknowledgements

For all his help and support, I would like to thank my thesis supervisor Dr. Mathias Schulze. During my years at the University of Waterloo, and later from afar, he did not only help me with any problems with respect to my dissertation, he also taught me what it takes to be a diligent researcher and instructor. I am grateful that I was able to benefit from his deep knowledge and understanding of the subject matter. Thank you for always treating my like a colleague and friend.

I would also like to thank everybody in the Germanic and Slavic Studies Department for the teaching, training, and administrative support they provided me and for giving my the chance of pursuing my graduate studies in Canada.

For their support and understanding in the months it took to complete the work, I would also like to thank my colleagues here at the University of Saskatchewan. A special thank you goes out to our chair, Richard Julien, who not only took the time proofread the entire manuscript, but also supplied me with abundant amounts of coffee and took me on the odd motorbike tours in order to keep me sane over the last year.

The software developers at the Technical Group at the Max Planck Institute for Psycholinguistics in Nijmegen, the Netherlands taught me what it takes to turn a n00b into a geek for which I will always be grateful.

For putting up with my quirks through all these years and lending me her love and moral support, I would like to thank my wife Christine. There were times when I thought that I would never finish this dissertation. Without her, I would never have.

Dedication

To Christine

Table of Contents

List of Figures

List of Tables

Nomenclature

ADDIE	Analysis, Design, Development, Implementation, Evaluation
AFF	Affix
AI	Artificial Intelligence
API	Application Programming Interface
BNC	British National Corpus
CALL	Computer Assisted Language Learning
CPU	Central Processing Unit
CSS	Cascading Style Sheets
DDL	Data Driven Learning
ESL	English as a Second Language
GPL	General Public License
GUI	Graphical User Interface
HCI	Human-Computer Interaction
HPSG	Head-Driven Phrase Structure Grammar

HTML	HyperText Markup Language
ICALL	Intelligent Computer Assisted Language Learning
KWIC	KeyWord In Context
MGG	Mainstream Generative Grammar
NLP	Natural Language Processing
POS	Part Of Speech
RISC	Reduced Instruction Set
SLA	Second Language Acquisition
SOAP	Simple Object Access Protocol
SQL	Standard Query Language
UG	Universal Grammar
UTF	Universal character code Transformation Format
XML	eXtended Markup Language

Chapter 1

Introduction

What can a software application look like that can potentially be used to help learners of a foreign language—and more specifically, learners of German as a foreign language—to extend their active and passive vocabulary, deepen their insight into the systematic rules that govern German word formation, and to improve their reading comprehension skills, and how does this sofware fit within current CALL applications, CALL theory and practice?

Can the computer serve as a tool to assist learners to achieve their goals by providing them with a range of features that are intended to help them work with a text in the target language of their choice?

Is it possible to develop an application with these capabilities that can be used in a classroom context, but that learners can also use independently?

Trying to address questions such as these which are at the centre of this dissertation falls into the realm of Computer Assisted Language Learning (CALL). One of the following chapters examines this discipline in detail. I will provide an overview of what CALL is, where it originated and how it might evolve. As CALL is a fairly young and diverse discipline, it is also necessary to look at the research paradigms that have been employed

by prominent researchers in this field.

With regard to my objectives, it is important to make some general remarks on CALL research. CALL is interdisciplinary by nature and it comes as no surprise that the approaches researchers have used reflect this diversity. CALL draws on the expertise of a wide variety of different disciplines, such as applied linguistics, computer science, second language acquisition (SLA) studies, psychology, sociology, philosophy, physics, mathematics, and many others. But it also draws on the methodologies used in these disciplines. CALL research can look — and has been looking — at how learners benefit from using a particular technology or a particular software, they can study whether a certain technology yields better learning outcomes than traditional language instruction, and similar issues.

There are researchers, however, who have been following a research paradigm (Levy, 1999) similar to the one I have employed for this dissertation. Working in this paradigm means not only to analyse and theorize, but also to develop new technologies. The method presented in chapter 4, in short, posits an iterative work flow that includes a needs analysis, the design and development of technology to address this need, its implementation and evaluation (Colpaert, 2004). At any given stage in this process, situations may arise that have implications for some of the other stages. At this point, the researcher decides whether to continue following the work flow or to address the situation, turning her attention to other affected stages. The work flow is very similar to the life cycle of software in the context of software engineering. A program is created to address a certain need or requirement. It is evaluated and will continue to be improved until it is eventually decommissioned, because it is no longer needed, or is replaced by another program that can fulfil the function more adequately, faster, or cheaper. The fact that the developmental paradigm is iterative in nature has some important consequences:

- During the design, or the development phase, developers may come across prob-

lems that make it necessary to consider reformulating goals that were developed in the analysis stage.

- During the design, or the development phase, developers may find technologies that they consider useful or interesting enough to warrant going back and include them in the goals.

- At the end of every evaluation phase, developers will return to the analysis stage and try to address the findings of the evaluation by introducing improvements.

- Work at a project is both time and labour intensive

- **Most important:** Evaluating the project at any particular time, before developmental efforts stop can only be done based on a developmental snapshot, representing the functionalities, abilities and shortcomings of the software at a particular time.

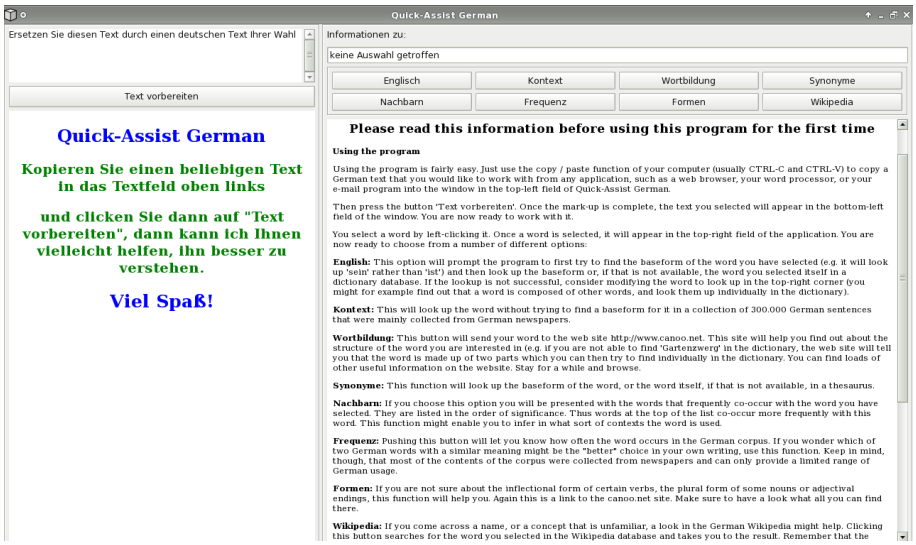


Figure 1.1: Startup Screen

QuickAssist, the program developed as part of this dissertation, addresses the questions outlined at the beginning of this introduction. In its current form it represents a

snapshot. It provides learners with the ability to import any German text of their choice that is available in electronic form and enables them to look up individual words in an English German lexicon, study these words in different contexts using a corpus of newspaper articles, provides synonyms, and information on word frequencies. If an internet connection is available, the program is also able to interface with the Canoo website (*canoo net*, last accessed: 13 September 2010) that provides a wider range of information such as the morphological analysis of words, their inflectional paradigms and grammar explanations. In order to find up-to-date information on German politics and culture, it is also possible to look up words in the German Wikipedia (*German Wikipedia*, last accessed: 13 September 2010). QuickAssist has a number of features that make it unique. It is platform independent, can be used with benefit by all but beginner German learners, does not use licensed resources, will be released under the General Public License (*General Public Licence*, last accessed: 13 September 2010) and is designed to be extended fairly easily (see chapter 4 for details). It was designed this way so that anybody interested in the program can use it and adapt it to their specific needs. The author hopes that this will be of benefit to all learners of German who would like to use this software and to other CALL developers who might be interested in the underlying source code.

This dissertation presents a number of salient findings regarding the usability of QuickAssist. Based on these findings, ways in which commercially available software could be improved are suggested, and directions for future research and development are discussed.

The software continues to improve based on the suggestions I receive from CALL colleagues, as well as from students and instructors using my software.

Before turning to CALL and to software design, the dissertation will review relevant areas of applied linguistics. The acquisition of new vocabulary and of word formation rules will be studied in some detail in chapter 2. Both of these areas are dealt with by the

discipline of second language acquisition studies (SLA). Some of the more recent models of how second language vocabulary is learned, stored, and what factors facilitate or hinder acquisition will be discussed. In addition, the theoretical underpinnings of morphology in general and German word formation in general will be considered. German, compared to English, has a richer morphology. Its use of nominalisations, and the fact that compounds in German are written together account for the fact that learners of German are faced with fairly complex wordforms early in their "career", such as the following:

(1.1) *Ein - wohn - er - melde - amt*
AFF - live - AFF - report - office
registration office for citizens and permanent residents

(1.2) *Führ - er - schein - prüf - ung*
control - AFF - certificate - examine - AFF
driving test

AFF, here, is used to denote a derivational affix.

While English has a similar rule to produce compounds, spelling conventions in English usually separate the individual constituents with whitespace. Learners of German, thus, have to learn early on not to be flustered by long words and to analyse them into smaller constituents in order to decode their meaning. The literature review will show that although these phenomena have been dealt with in linguistic accounts of German morphology and word formation, there has been little effort to make these insights available to learners by introducing this area of language to curricula or dealing with them in the more prominent textbooks available for German as a foreign language (henceforth abbreviated DaF — Deutsch als Fremdsprache — I am not aware of an equivalent English abbreviation that is used in literature consistently). Chapter 2 concludes that German morphology is governed by rules, just as German syntax is, and that these rules can and should be learned. While this might seem obvious, research has shown that there is a shortage of adequate teaching materials devoted to instructing learners about these rules.

In chapter 3 the focus shifts to CALL. Because of the discrete nature of vocabulary items, they have been of considerable interest to CALL developers. It is relatively easy to develop routines that present vocabulary items to learners and at a later stage ask them to reproduce them. The range of possible correct translations for a certain word are limited. It will be argued, however, that looking at vocabulary in a textual context and outside of one-to-one L1-L2 pairings is a non-trivial task.

CALL software has been dealing with vocabulary in a variety of ways. Vocabulary trainers have been developed in order to provide vocabulary drills of various kinds, such as translation exercises, fill-in-the-gap style exercises, and so forth. Some programs provide users with pictures to increase the saliency of new vocabulary items, or acoustic representations in order to enable learners to practice the pronunciation of new words. Not long after the first language corpora were created, researchers started to experiment with them in the context of foreign language teaching. Data Driven Learning (DDL) Johns (1991) refers to the use of concordancers in foreign language teaching. Students are provided with lists of sentences with a specific word in different contexts: keyword in context (KWIC) lists. Studying these lists will help them, that is the hypothesis, to infer the meaning of the word and get an idea of its semantic scope and pragmatic usages. Creating and working with language corpora and concordancers falls into the domain of natural language processing (NLP). I will argue that integrating corpora, concordancers and other NLP applications into CALL program can enrich them and provide useful tools for language learners. The QuickAssist user study that is described in chapter 6 is one case in point.

It is necessary to discuss what sort of technology is available in the field of NLP and how it can potentially be used in CALL applications in order to be able to assess whether currently available commercial CALL software can be considered state of the art. If this is not the case, the obvious question is which of these technologies can be used in CALL

and for what purpose. Some of the more prominent technologies provided by NLP, their potential benefits and pitfalls of their use will be presented. Applying NLP and Artificial Intelligence (AI) technologies to CALL is commonly referred to as ICALL (Intelligent CALL). It will be argued that, although many of the technologies have been working fairly robustly for a number of years, commercial software publishers have been hesitant to adopt them for CALL applications. Possible reasons for this will be disussed.

The perspective I have chosen, is to adopt NLP with its current capabilities for CALL applications, which may radically change the traditional roles of language learners, teachers and the computer in the CALL context. The role of the computer in CALL has traditionally been described with the help of dichotomies. It can either act as a tool or tutor (using Levi's 1997 terminology). These terms are roughly equivalent to other dichotomies, such as magister/pedagogue, used by Higgins (1988). Details will be discussed in section 3.2. However, one way of interpreting these dichotomies is to asses the degree of control assigned to the computer and the degree of freedom the user is given in deciding what to do herself. The dichotomy, from this angle, becomes a continuum. I will argue that dedicated commercial CALL software can be located at the tutor/magister end of this continuum.

The reason for failing to make the shift to student centred CALL cannot be accounted for by technical limitations. On the contrary, if the power to make decisions about the learning process is shifted from the computer to the learner, the necessity to have it act omniscient and human-like disappears. The student in the student centred CALL context is informed about what the computer is capable of doing and has been made aware of its shortcomings, assumes responsibility for her learning process and can rely on the support of human instructors to help her reach this level of independence.

Taking the importance of vocabulary and the knowledge of word formation rules as a point of departure and adding student centred CALL as a desideratum, chapter 4 discusses

the development of QuickAssist. The development was guided by the design principles for CALL applications, laid out in Colpaert (2004) which have evolved into a de facto standard for CALL software development. In addition, I will also return to the question of the role software development plays from a CALL research perspective. This chapter is also the most technical in nature.

It was the most challenging (and most time consuming) part of this dissertation to become sufficiently computer literate to master the programming skills required to develop QuickAssist. Time was spent learning various programming languages and testing different database technologies and GUI (Graphical User Interface) tool kits to find the most suitable ones for the task. Section 4.4 provides a detailed discussion of this aspect.

The architecture of QuickAssist is described in section 4.6. In section 4.2.1 some issues connected to CALL programs and programming in general are discussed. Drawing on Wood (2008), the problem of commercial closed code licences for CALL, NLP software and other software used in education and research is explained. It will be concluded that in order to facilitate advance in any field of computer programming it is necessary to distribute software together with its source code. A brief discussion of legal rights and obligations, authorship issues in successful open source initiatives, as well as platform independence will conclude this chapter.

Chapter 6 presents the methodology used to evaluate the software created as part of this project. It contains information on criteria commonly used to evaluate software and discuss the standards that exist for the evaluation of educational software. There are suitable methods to evaluate the usability of software and to gain an understanding of its abilities and shortcomings. Qualitative methods have proved to provide detailed insights large scale quantitative studies are not able to produce.

User walkthroughs were used to evaluate QuickAssist. This qualitative method of software evaluation has been used and described by Hémard (1999) and others. In the

QuickAssist user study, users were able to relate their experiences to the interviewer in far more detail than one could hope to achieve with any software evaluation form. In addition, they had the freedom and were encouraged to comment on any other aspects of the software. My test group consisted of four upper intermediate to advanced learners of German and three instructors of German. While the learners were given a task to complete in a set period of time, and asked to comment on what strategies they used to complete the task, the instructors were asked to experiment with the software and then asked to evaluate it by using questions adapted from a standard software evaluation form. Further, the learners were asked to use the software for one month and report on their experiences in a final interview.

Chapter 6 will provide detailed information on the participants of the study and will summarize the results, concentrating on the question of the usability of the program for learners of German at different levels and on its potential benefits for instructors. It will also report on what users perceived as benefits and problems of the software.

In chapter 7, I will argue that QuickAssist is an application that can be successfully used by the intended audience and for its intended purpose. While similar programs exist, QuickAssist is a project that is based entirely on reusable resources, licensed under the GPL, and offers a larger pool of options to learners than Glosser RuG(Dokter et al., 1998), which is the most comparable program. It is largely platform independent and can be adapted to specific settings and extended to handle other languages.

The qualitative study yielded a number of interesting results. These, of course, need to be verified in follow-up studies, possibly quantitative ones, using control groups to establish in which respects the learning progress of users of QuickAssist differs from students not using the program.

As pointed out, the development of software is not completed until it ceases to be used. The input received by the participants of the study combined with that of people

who attended my public demonstrations have provided a plethora of development tasks for years to come. The dissertation will culminate by highlighting some of the features that I want to add in the future, e.g., the option to create automatic exercises to practise new vocabulary.

Chapter 2

Theoretical background

2.1 Overview

In order to motivate the importance of a CALL tool like QuickAssist, it is necessary to consider a number of different things.

Since QuickAssist is intended to help learners of German, it is concerned with human language. In this chapter, I will look at the areas of linguistics that play an important role with regard to the intended use of QuickAssist. This program is concerned with words, their structure and their meaning.

The study of the structure of words, morphology, will be the topic in section 2.2. The section is not intended as a general overview. Since the computational representation of language is an important issue for computational linguistics, which is discussed in Chapter 3, I am concentrating on how linguistic theory can help to inform computational linguistics.

In section 2.3, the processes used to form German words and their underlying rules will be considered in some detail. This section is intended to show that while German

derivational morphology might appear to be a complex system, it is rule governed. Learners can and ought to become familiar with this system.

Section 2.4 will start out by discussing the area of second language acquisition studies (SLA) in general and then concentrate on the area of vocabulary acquisition. While vocabulary is usually only associated with the meaning of words, we will see that this is only one aspect. Knowing a word involves many different kinds of knowledges and the ability to analyse it and identify its individual constituents, their form and function, is another important component of vocabulary knowledge. I will also briefly deal with methods of acquiring and extending the vocabulary range in a foreign language and discuss what contribution extensive reading can make.

Section 2.5 concludes this chapter and is intended to contrast theoretical concepts and empirical evidence from SLA with the reality of teaching or learning a foreign language such as German. The conclusions are that the treatment of vocabulary and especially word formation should be more extensive than it currently is, according to research literature. How the computer can be used to provide additional opportunities for learners to extend their vocabulary and knowledge of word formation processes will be the concern of the remainder of this dissertation.

2.2 Morphology

2.2.1 What is morphology?

In this text I assume that morphology is part of the grammar of a language. This is probably not contentious. It is a matter of debate, however, what grammar is and what it comprises. Here it will be assumed that a grammar of a language are the rules that underlie the use of this language. It can be further subdivided into syntax, morphology, and

lexicology. In some cases, authors have chosen to include other areas such as phonology, semantics and pragmatics, too (for a discussion see Simmler, 1998). While syntax is mainly concerned with the order of elements in sentences and/or utterances, morphology is concerned with the structure of words. Lexicology in return deals with the meaning of words and constructions. It attempts to describe the nature of the lexicon, what it contains and how it interfaces with the syntax and morphology (cf. Singleton, 2000). The dissertation will be concerned with words and their meanings later, when I turn to the area of vocabulary acquisition.

Morphology can be further subdivided into inflectional and derivational morphology. The majority of literature considers compositional morphology a third branch of morphology. If viewed from a language acquisition perspective, composition, reduplication and many other phenomena can be considered to fulfil the same function, i.e. forming a new word by combining existing words, roots and affixes, or at least using a word belonging to one category in a manner as if it belongs to another category (conversion). The result is always the same, from a functional point of view. Therefore, here, derivational morphology will be considered to include these phenomena. Derivational processes are governed by rules similar to the rules that govern the order of elements in a sentence. These rules are equally complex as syntactic rules, but in foreign language learning they can be studied and are worthwhile learning.

Inflectional morphology can be considered to form the interface between syntax and morphology. It is also sometimes referred to as morphosyntax (Culicover, 2009), although this usually implies a syntactic perspective on inflection. Inflection is the morphological marking of grammatical functions on words. Inflection succeeds derivation:

(2.1) *Schreib - tisch - täter - s*
write - table - perpetrator - 's
white collar criminal's

In this example, a word that is the result of two derivational processes is finally marked as a genitive. In German, nouns, verbs, adjectives, determiners and pronouns are inflecting word classes. I will argue later that inflectional morphology is being dealt with sufficiently in literature and forms an important part in the majority of DaF classes. In this project, I mainly will be concerned with derivational morphology.

2.2.2 The need of formal accounts of morphology

Syntax and grammar in general can be dealt with in a variety of ways. There is number of books on the market, for example on "the grammar of English". The majority of them usually turn out to be stylistic guides. They are referred to as prescriptive grammars and address readers that aim at improving their language use in order to more closely approximate a variety of English that is preferred in professional discourse and often considered superior to other varieties. These "grammars" are usually lists of dos and don'ts along the lines of "do not use ain't", "avoid the passive voice", etc.

Descriptive grammars, on the other hand, attempt to describe what the rules are that underlie actual language use. This provides a linguistic account of the state of affairs, rather than the attempt to change this state of affairs. It should be noted, however, that even descriptive accounts, necessarily, have to restrict themselves to one variety or a small number of language varieties and will always have to exclude others, although this is done for different reasons.

While descriptive grammars usually try to be accurate, in most cases they are not exhaustive. While they are able to address most questions and problems, people may have regarding grammar, many details are left out, or are not described in a way that can be used to implement this descriptive account as a working computational model of a language in a straight forward way.

The situation for morphology is fairly similar. As regards German, the only accounts for morphology available are descriptive. To my knowledge there is only one publication that tries to formalise the rules of German word formation (Motsch, 2004), and as I will argue, it did not do so very successfully.

If we were able to describe German morphology in a formal way, a computational model of German word formation could be developed. Moreover, if the formal account is accurate and psychologically plausible, it could form the basis for teaching material geared toward teaching word formation. This issue will be revisited at the end of this section.

The linguistic discipline that is most concerned with formal accounts of natural language is generative grammar. The following section will take a look at this discipline and more specifically at the status of morphology in different frameworks.

2.2.3 Generative grammar - formal accounts of morphology

Generative grammar, no matter in what flavour it comes, attempts to give a formal account of how sound patterns in a language are related to meaning by syntactic and morphological rules and constraints. It is generative in two respects. On the one hand, it tries to establish rules that can generate grammatical sentences. On the other hand, it has been referred to as generative because of these rules that the various theoretical frameworks postulate, or generate (Jackendoff, 2003). By attempting to give an accurate, exhaustive and (at times at least) simple account of how this is achieved, generative descriptions lend themselves particularly well to form the basis of computational models used to generate or process natural language. They do because they are formal in the mathematical sense and relatively easy to implement as programming algorithms.

Taking a look at SLA literature that is concerned with theoretical linguistics in general,

and generative grammar in particular, one gets the impression that one of the main points of interest seems to be the concept of Universal Grammar (see for example: Cook & Newson (2007)). As Jackendoff (2003) points out, nowadays the main concern of SLA researchers in this regard seems to be to refute the notion of Universal Grammar (UG), which Jackendoff admits may very well be due to what has come to be referred to as "the linguistics wars" (Harris, 1995) in which Jackendoff himself was one of the key proponents of the Chomskian framework.

I will not have much to say about the plausibility of the UG hypothesis, nor on the question what role it plays in second language acquisition, if one was to accept it as a plausible concept. I will also have little to say about the importance of generative grammar for SLA as a whole, but I would concede with Cook (2003) that without the Chomskian concept of grammar as a set of rules and constraints that enables us to produce and understand all the possible utterances of a language, even if we have never heard or read them before, many concepts that are integral parts of many SLA theories would not exist. This ability can not be accounted for by earlier models of language acquisition, such as the behaviourist model which conceptualises language acquisition as a series of stimulus response incidents. Without the notion of rules licensing possible sentences, the concept of interlanguage (see for example Gass & Selinker, 2008), a learner grammar, could not have been conceived of.

Pertinent to the subject of this text is in how far the theoretical models of a particular flavour of generative grammar can be directly applied to the area of NLP. In the following discussion, I will concentrate mainly on the area of word formation and inflection, subjects that are usually dealt with by linguistic morphology (Booij, 2005). This, however, has not always been the case.

In the beginning, that is with the publication of Chomsky (1957), generative grammar had no need for a morphological component. The syntactic component was responsible

to generate both a syntactic structure and also filled the slots this structure had for words from a minimal lexicon that only contained roots and affixes. Neither did this model have any need for a semantic component. Jackendoff (2003) calls this syntactocentrism, and although the relative importance of the lexicon within Chomsky's framework increased over time, the emphasis is still very much placed on syntax in its latest incarnation, the minimalist program (Chomsky, 1995). These lexical insertion rules, as most of the other rules of the early approaches, were very powerful. They did not only insert the roots and appropriate affixes, but could also modify the root (such as is in the case of irregular plural forms like *goose/ geese* for example). As the roots and affixes were stored without any additional information, the insertion rules where not able to infer whether certain roots could be used to form a member of a particular word category. This led to some theoretical models in which all words were thought to be derived of verbs. Special rules would then transform these forms and adapt them to the context they were to occur in. Thus if the noun *runner* was required, the verb *run* would be retrieved, it would be affixed with the suffix *-er* and the result would be inserted. Leaving aside the question of how the extra *-n-* is inserted, even more importantly, this made it necessary for the model to claim the existence of verbs like *to king*, *to queen*,... in order to account for the respective nouns.

It soon became apparent that a minimal lexicon and a set of primitive lexical insertion rules were not able to account for the vast majority of sentences. Over time, the lexicon necessary for the description of syntax grew in size, the entries got more complex in that information about their category, contexts of appearance, restrictions, etc. were added. The morphological component slowly made its way out of the syntactic component where it was first conceptualised as a "black box" in which "magic happened." That is to say that the need for such a component was acknowledged, its in- and output stipulated by the theory, but its inner workings were not discussed in any detail. It was later alternatively

assumed to be an individual component, or part of the lexicon, and provided with various degrees of interfacing with the syntactic component. For a more complete account of early generative morphology see Scalise (1984).

Later generative approaches to Morphology conceptualised affixing as happening in a certain order, along a series of strata. The number of strata postulated varies widely between different theories. Scalise (1984) uses up to six for his account of Italian affixes, with a whole stratum responsible only for diminutives, while Katamba (1994) uses two strata for his account of English morphology. While it has been shown that the strata hypothesis has some problems (see Giegerich (1999) for details), it still offers a detailed account of the order in which affixes are applied to a base in derivational and inflectional processes.

What Culicover & Jackendoff (2005) call "alternative generative approaches" such as Lexical-Functional Grammar (Bresnan, 2001), and Head-driven Phrase Structure Grammar (Pollard & Sag, 1987, 1994; Sag et al., 2003) differ from MGG (Mainstream Generative Grammar) (Culicover & Jackendoff, 2005). They are non-derivational, that means they try to do away with the need of concepts such as D-structure and invisible constituents. Introducing simplicity on the syntactic level, on the other hand, increases complexity elsewhere, and this is at the level of the lexicon, which in all of these frameworks is far richer than in the Chomskian ones. Lexical entries contain detailed information on idiosyncratic features of each entry, such as its meaning, what other lexical entries it subcategorises, as well as information on morphological features. While this in itself is not a move toward an independent morphological component, it acknowledges the fact that there is regularity on a sub-syntactic level. Jackendoff (2003) rightly points out that the way morphology is dealt with within HPSG is not able to account for online production and analysis of complex word forms. On the other hand, works like Riehemann (1993) and Riehemann (1998) show that this framework is able to describe German derivational

morphology in a very detailed manner. Additionally, it has been argued (Pollard, 1988) that Categorial Grammar (Hoeksma, 1985) and HPSG are largely compatible and there are computational models of German morphology (e.g., Schulze, 2001) that use both HPSG and Categorial Grammar.

Apart from the various generative frameworks discussed in theoretical linguistics, other models have evolved in the area of computational linguistics that certainly are to some extent inspired by the models mentioned above, on the other hand they were developed with the specific problem of implementing them as a computational model in mind. As regards the area of morphology, the seminal work here was Koskenniemi (1983). A more detailed discussion of this and other NLP models will follow in chapter 3.

Incidentally, it should be noted that the models of the morphological component and the lexicon, as well as the models about how these components interact and information is accessed seem to be evolving in parallel with computer technology. For example: the first computers were unwieldy, operated fairly slowly and had very little memory. Also, programs proceeded sequentially, one step after another. Likewise, the first models of generative grammar such as Chomsky (1957) or Chomsky (1965) were based on the idea that the number of elements needed to form words and sentences had to be as small as possible. Although little is said about performance, it is clear that processing was conceived of as happening iteratively in unidirectional order, in much the same way in which finite state automata function. The merge and transform rules (although they had different names then) function in the same way as do move and copy operations in low level computer languages. The same is true for early accounts of lexical processing (see Singleton, 2000) which assumed that the mind processes the input in a linear, unidirectional order and attempted to account for the processing without having to postulate a large storage space.

Looking at current models of generative Grammar such as Chomsky (1995); Bresnan

(2001); Pollard & Sag (1987); Sag et al. (2003); Culicover & Jackendoff (2005), while these authors claim that their accounts of grammar are psychologically more plausible than their predecessors' (apart from Chomsky: notwithstanding his demands for psychological plausibility in his early works, he has meanwhile changed his view, creating a framework postulating invisible structures and empty elements which bear no resemblance to what one might call psychologically possible), they all have in common that the size of their lexicon has grown over time, just as the size of memory of the average computer has grown over the years. You could even argue that trying to simplify and reduce the set of rules a model has to postulate in order to operate adequately finds its resemblance in the development of the first RISC (reduced instruction set) processors. These processors that were first developed at the end of the 1980s were able to increase the processing power and speed by several magnitudes, because they used a dramatically reduced set of internal commands compared to common microprocessors.

More recent models of lexical processing (see Singleton, 2000) assume parallel processing which in return finds its computational equivalent in multi processor computers, distributed programming and fitting modern microprocessors with multiple CPUs. This, I think, shows that although many linguists claim that our brain does not function like a computer — and this assumption seems to be born out by the advances of neurolinguistic research — the computer–brain metaphor still seems to hold.

2.2.4 Available accounts of German morphology

Returning to morphology: after this brief overview of how morphology is conceived of in different generative frameworks, I am turning now to the more recent accounts of German morphology and word formation that attempt to be comprehensive. In doing so, I will show that the accounts of German morphology available do not meet the demands of software developers and of DaF practitioners alike. The latter group comprises instructors

as well textbook authors and curriculum designers.

Motsch (2004) is the only work I have consulted that attempts a description from within a particular linguistic framework which, in this case, is a Chomskian one in the widest sense. Motsch gives a detailed account of all productive, semi-productive, and to some extent unproductive word formation processes in German, developing a special notation to arrive at a formal description of grammatical and semantic features of all word formation processes. This is probably one reason why it has not received wider attention in Germanic linguistics, where Chomskian linguistics does not play an important role, at least not to the same extent it does in North America. The choice of framework and the idiosyncratic system used for the description of word formation processes, I would argue, has rendered it fairly inaccessible for the majority of German linguists with an interest in word formation. Moreover, I am not aware of any attempts to use Motsch's analysis as the basis for a computational model.

If there is any difference between German linguistics and North American linguistics, it is probably the fact that the majority of linguists in Germany are fairly agnostic when it comes to the question of frameworks they are using (this, I might add, holds true only to a lesser extent in the area of computational linguistics). Römer (2006, p. X) cites Klein (2004) who writes:

> Weg von den engen ‚frameworks' und ihren idiosynkratischen Begrifflichkeiten

which translates, "let's do away with narrow perspectives and their idiosyncratic descriptions". This will become even clearer after taking a closer look at the remaining accounts of morphology and word formation.

The classic work in this area is Fleischer & Barz (2007) which was first published in 1983 (which in turn is partly based on a publication dating back to 1969) and has un-

dergone a number of revisions and republications. Without subscribing to any particular linguistic school of thought, the authors give a descriptive account of word formation rules and affixes. They are concentrating on contemporary standard German, but provide examples from various periods of German literature and other texts in order to illustrate word formation patterns that are no longer transparent. Although the treatment of different affixes is very detailed, the fact that makes it difficult to use, either in an educational setting or for the development of NLP software, is that descriptions are in prose form, the language used contains elements of modality and similar hedging devices that make it hard to formalise the analyses of the authors or infer from them a set of rules in a straight forward way that German learners could use.

Simmler (1998) attempts to to give a comprehensive description of German morphology, including both inflectional and derivational morphology. After defining morphology as a sub-discipline of grammar, but separate from syntax and lexicology, he develops a complex system to describe morphology, stipulating six different morpheme types which in turn can be further divided into different classes. In the chapters dedicated to word formation, he considers, just like Fleischer & Barz (2007), word formation processes for the lexical categories (noun, adjective, verb) and adverbs, but also looks at pronouns, conjunctions and prepositions. Occasionally, he provides empirical evidence such as frequency counts and other productivity measures, which I would argue, are absolutely necessary to determine in how far a certain word formation process can be considered productive or unproductive.

It is of theoretical interest which word formation processes were once productive and how the effects of them can still be observed to some extent in contemporary German. Learners and/or their instructors, on the other hand, need to know about what the productive and, potentially, semi-productive word formation processes are in order to analyse unknown complex words, or creatively form new words that speakers of German are likely

able to decode. By establishing, in an empirical way, what processes are no longer productive, the economy of learning can be enhanced (e.g., learners will not have to spend any time on analysing words that are the result of unproductive word formation rules). Neither the structure, nor the original meaning of the individual components will help them determine the contemporary meaning of the word, nor will it help them in analysing other words that are equally opaque. It has been shown by many authors (e.g., Baayen & Lieber, 1991; Plag, 1999; Wood, 2002) that measures such as Zipf's rule which is based on the frequency of hapax legomena in language corpora can be used in order to provide empirical evidence for or against the productivity of a word formation process.

Although Simmler (1998) provides some empirical evidence, he does so only occasionally. It is not possible to access the degree of productivity of any given word formation process in a quick and straight forward way with either of these references. That together with the fact that affixes and word formation processes are discussed in prose style rather than defined in a concise, reference style manner, makes this text an unsuitable tool for learners, teachers, and NLP developers. This is not to say that it should be. The intended audience of both texts probably never included DaF practitioners or CALL developers. It demonstrates, nevertheless, that there is no literature available for these audiences. Both Simmler (1998) and Fleischer & Barz (2007) are scholarly works in their own right, but they lack the exhaustiveness and structure necessary to make them a suitable tool for CALL and NLP developers and prove fairly inaccessible for DaF practitioners for the same reasons. It may be the paucity of suitable references that is to blame, in part, for the dramatic shortage of teaching material that is devoted to the teaching of word formation rules in German.

Riehemann (1993) has shown that it is possible to give a formal and exhaustive account of German word formation processes by demonstrating it for *-bar*, a suffix that derives adjectives from verbs. By introducing lexical hierarchies, she is able to give a

complex and detailed account of this word formation process, syntactic and semantic features of the bases participating in it and syntactic and semantic features of the derivatives. On the other hand, her work displays that it would be a tremendous undertaking to do the same for all other productive and semi-productive affixes that exist in German. Fleischer & Barz (2007) list about 450 affixes in their register and arguably their list is not exhaustive. Considering the resources that would be required to realise this task, it can be safely assumed that it will not be undertaken, let alone completed any time soon.

It might then seem as if German morphology is too complex to be adequately represented in a computational model. We will return to the issue of adequate computational models in chapter 3.

It is interesting to note that there have been relatively few monographs that have been published on German morphology. The standard introduction for many years was Bergenholtz & Mugdan (1979). A widely used introduction to German Grammar for students of linguistics used to be and still is Eisenberg (1985). Although the title promises that the book provides an overview of all the elements of German grammar, the book only covers word categories and syntax. Even though it was republished a number of times, this did not change until the fourth edition when it was decided to turn it into multiple volumes, the first of which bears the subtitle "Das Wort" (cf. Eisenberg, 1998), while the second one is called "Der Satz". This is even more surprising since a lot of research was conducted on German morphology, indeed, the entire field of natural morphology can be said to be dominated by German speaking linguists such as Wurzel, Dressler, and Mayerthaler.

Before turning to German word formation then, we can note that German morphology, overall is a well researched field, but that the standard works concentrating on the derivational component are limited in their usability, both by software developers and DaF practitioners.

2.3 German word formation

2.3.1 German word formation rules

To express complex concepts, i.e. concepts that can not be represented by simplex items, German speakers/writers can resort to one of two options. Either they can use a phrasal or sentential expression to describe the concept, or they can use a complex word to refer to it. The latter method makes use of a set of systematic rules that German speaker/writers use to form complex words and that readers/listeners can use to analyse complex forms in order to determine their individual constituents and establish their meaning.

The number of truly simplex items in German is relatively small. Most words consist of more than a single morpheme. Many of these complex items are probably stored as single, unanalysed items in the lexicon of native speakers. These are used with such frequency that it is not necessary to synthesize or analyse them again and again. Processing this word over and over again would seem less efficient than simply storing it in its entirety. This is why most cognitive theories assume that the lexicon contains complex items as well as simple items. Most of these complex items, however, could be analysed by native speakers. They would be able to identify the individual constituents and would be able to talk about their semantic properties.

Thus, in the following example, German native speakers would be able to say that *Briefträger* consists of two constituents: *Brief* and *Träger*, that *Brief* is a letter and that *Träger* means carrier. They are also aware of the fact that *Träger* is derived from the verb *tragen* and that the derivational affix *-er* can attach to verb bases in order to form a noun denoting a (male) person who carries out the action denoted by the verb. Ultimately then, the word is made up of three constituents.

(2.2) *Brief* - *träg* - *er*
letter - carry - AFF

postman

On the other hand, there are a number of words in German that were once formed using productive word formation processes, but today these processes are no longer productive and the words remain unanalysed. This is especially true for foreign loan words, which German speakers adapted as monolithic units, but also for word forms that contain elements that modern speakers do not recognize as morphemes, because they ceased to exist in any other contexts. For example, while

(2.3) Helikopter

is made up of two Greek morphemes, *helico-*, from *helix* (spiral), and *pter* from *pteron* (wing), the structure of the word for most speakers of German is opaque and they will analyze the word as a simplex item. The affix *-ig*, on the other hand, is a moderately productive German affix turning nouns into adjectives that have the meaning: *having the property of*:

(2.4) salzig (salty)

(2.5) farbig (colourful)

(2.6) schusselig (clumsy, Schussel means scatterbrain)

(2.7) ledig

Nevertheless, most German speakers are not aware of the fact that *led* in Old High German has the meaning: *part*. Today it is no longer in use and only exists in this adjective meaning *single*. Therefore, we can assume that the word is stored as a simplex item in the lexicon.

German, like other natural languages, undergoes constant change. New elements find their way into the language, while others stop being used. These processes are gradual,

and while most modern linguistic theories would like to treat language synchronically, only the study of processes like lexicalization and grammaticalization and other mechanisms of language change will be able to present a clear, diachronic picture of these changes (Hopper & Traugott, 2003; Lehmann, 1995).

While not all word formation processes are active, at any given point in time, a number of them are available and are used to form new words for new concepts, or –to be creative– new words for old concepts. Without the ability to use these rules, people would not be able to analyse words such as the following, or to construct new words themselves using systematic rules.

(2.8) *dis - öko - log - isch*
dis - eco - log - ical
not ecological

Rule: opposites of adjectives with a Latinate base can be derived with *dis-*

(2.9) *erz - bereit*
ultra - prepared
ultra prepared

Rule: emphasis can be added to many adjectives describing character traits by prefixing *erz-*

(2.10) *Plärr - minator*
shriek - minator
someone who shrieks a lot

Ad-hoc formation formed as an analogy to *terminator – governator*

Derivational morphology is the linguistic subfield that studies word formation processes and the following section should serve as an overview of the most important processes that exist in modern German. The terminology I am using is adapted from Donalies

(2007). This does not indicate that this is a universally accepted set of terms, however, it is a fairly recent one and relatively concise compared to others (e.g., Fleischer & Barz, 2007).

2.3.2 Units of word formation

In order to discuss the rules of German word formation, it is necessary to define how to classify the elements that are used in word formation processes. Following Donalies (2007), we use the following hierarchy:

Words in the current context are all word forms that can stand by themselves. E.g., while *Kugel*(ball) in *Kugelschreiber*(ballpen) can be found by itself, *-er* can not.

Phrases are sometimes used to create new words: while *malte den Teufel an die Wand* in Hans malte den Teufel an die Wand (Hans made us expect the worst) is a verb phrase, in *Das Teufel-an-die-Wand Malen ging allen auf die Nerven* the entire phrase is used as a noun, which German writers often indicate by hyphenating the entire phrase.

Letters: individual letters can form parts of a word. In *x-mal, T-Träger, B-Betrieb* the letters *x, T, B* are not abbreviations of other words, they are meaningful by themselves: *x* is used as a mathematical variable to indicate that something is perceived to have occured a large number of times. The *T* refers to the shape of the letter when used together with *Träger* to denote a cantilever that has a particular shape, and the *B* refers to a certain operation mode of semi-conductors, as opposed to the a, ab, and c mode.

Confixes in contrast to words cannot stand by themselves. They only occur in bound contexts. Examples in German are *therm, geo, bio, techno*. Contrary to affixes (see

below), they can combine with other confixes to form a word. In other studies, they are usually considered bound roots, which is less accurate, because this would mean that we have to analyze *log* as an affix in *geologisch* and as a base in *Logistik*. Most confixes are of a classical nature, but are productively used in neoclassical word formations (Lüdeling et al., 2002).

Affixes, just like confixes, can only occur with other elements and not by themselves. They can not, in contrast to confixes, combine with other affixes. German uses prefixes that attach to the front of a base, e.g., *vor-* in *vorschlagen* and suffixes, that attach to the end of a base, e.g., *-schaft* in *Freundschaft*. Whether it is necessary or useful to posit circumfixes for German is a matter of debate in German linguistics. Simmler (1998) provides an overview of the discussion and offers an alternative analysis. Schulze (2001) considers it a useful construct especially for a formal account of German morphology. Following this analysis, *Gebirge*, consists of the circumfix *Ge..e* and the base *berg* that undergoes a stem vowel change during affixation. It is equally debatable if German makes use of interfixes, or whether an alternative analysis of cases such as *-an-* in *Republikaner* is more adequate. An alternative, here, would be to analyse *-aner* as an allomorph of a morpheme that derives nouns from nouns and has the meaning *believing in* or *follower of* and has a number of different allomorphs. Examples include: *Demokrat, Hegelianer, Idealist, Romantiker*. A similar argument can be made for stem vowel changes that frequently occur in German and that can be interpreted as infixes or allomorphs respectively. E.g., *Männ - er*(men) can be analysed as an allomorph of*Mann*(man) and a plural suffix, or as*Mann* plus a plural affix, plus an infix that is responsible for the vowel change.

Donalies (2007) argues that affixes can be subdivided into four groups using two distinct features. The first feature is [+/-] transposing and indicates whether an affix

is used to change the syntactic category of the base it attaches to, the other one is [+/-] categorial meaning change and indicates whether the affixation results in a change of the semantic domain of the base. The resulting four types are:

- + : The affix leads to a change in the syntactic category while the semantics remain unchanged. E.g., *Faulheit* turns the adjective *faul* (lazy) into a noun, but both words refer to a specific property, so there is no semantic change. It could be argued that an affix changing the part of speech of the base it attaches to is to alter the lexical rather than the syntactic category. I will follow Donalies (2007) here. Changing the part of speech has an impact on the syntactic scope of the resulting word form. The main reason for forming the word *Faulheit* is not to change its meaning, nor to use a different inflectional paradigm with it, but primarily to use it in a different syntactic context.

+ + : The affix brings about both changes in terms of syntax and semantics. E.g., *Sensibelchen*(wallflower) turns the adjective *sensibel* into a noun, the resulting compound refers to a person now, not a property.

- - : The affix does not bring about any syntactic or semantic changes. E.g., *Häuslein* is the result of applying the diminutive affix *-lein* to the base *Haus*. The result is still a noun and refers to an entity. The affix does merely add the additional information that the entity referred to is perceived as small or quaint.

- + : While the syntactic category remains unchanged, the semantic domain changes. E.g.: adding the suffix *-heit* to the base *Gott* does not change its syntactic category. Both forms have the syntactic scope of German nouns. The semantics, on the other hand, are affected. While the base refers to a concrete concept, *-heit* indicates either an abstract concept: some form of god, or adds plurality to the feature set of the base: *Mensch*(one human), *Men-*

schheit(humanity, all humans).

While these categories are certainly not without problems, I think that they provide useful instruments to describe the functions of derivational affixes.

Unique units is Donalies' term for what linguists usually call cranberry morphemes (Carstairs-McCarthy, 2002). The term is used to refer to morphemes that from a synchronous perspective have no more independent meaning and only occur in a single context. While *Him-* in *Himbeere*(raspberry) goes back to the Middle High German word *hinde* (female deer), this is no longer transparent to most modern speakers of German who would either not be able to assign a meaning to *Him-*, or would possibly see a relationship to *Himmel* (sky), especially since this would be in opposition to *Erd-*(ground) in *Erdbeere*(strawberry). Reanalyses of this kind are considered folk etymologies.

Fugenelemente (linking morphemes (Glück, 1993)) are elements that are used in between other elements in compounds. Examples are *s* in *Abfahrtszeit, Staatsmacht, Kindheitstraum*, *o* in *germanophil, Thermometer*, or *i* in *toxigen*. While studies exist that aim at explaining regularities for these elements, there are only a few derivational affixes that *s* attaches to regularly:

- -heit: Freiheitskämpfer
- -ion: Flexionsklasse
- -ität: Identitätskrise
- -keit: Höflichkeitsform
- -schaft: Landschaftsmaler
- -ung: ahnungslos

There are two opposing views with regard to these elements. One position is that only non-inflected bases participate in word formation. From this perspective, everything between base and affix is considered a linking morpheme. Thus, *n* in *Leichentuch*(shroud) would have to be analyzed as a linking morpheme. The second position is that only those elements that do not fit into the inflectional paradigms of the base be considered linking morphemes. As *n* in *Leichentuch* is a plural morpheme that attaches to the base Leiche(corpse), it is no linking morpheme. While the use of the plural form seems to be semantically motivated in cases such as *Blumenvase* (a vase for flowers), or *Bücherkiste* (a box for books), this seems less so for *Leichentuch*, or *Mäusefalle*(mouse trap), although one could still argue that a mouse trap is used to catch mice and a shroud is used to cover corpses, even if both serve their purpose for one entity at a time. While many compounds can then be analysed as containing plural bases or genitive bases that can be interpreted as genitivus subjectivus, or genitivus objectivus, there seems to be no regularity in this regard. Donalies (2007) points out that while *Bücherkiste* (book(s) box), containing the plural form *Bücher* is a box for books, *Buchladen* (book store), a place where one also expects to find a lot of books, only contains the singular form.

Linguists have also claimed that the elements are also phonetically motivated, i.e., they facilitate pronunciation of the compound. According to Donalies (2007) there is little ground for this. There are a number of competing forms such as *Fabrikgebäude* which is mainly used in Germany and *Fabriksgebäude* which is used in Austria. This, by the way, is also the reason why translating *Fugenelement* with epenthesis would be misleading, as an epenthesis is phonetically motivated. I am using linking morpheme here, although this of course raises the question about the meaning of the morpheme if one follows the traditional definition that a morpheme is the smallest meaningful unit Bauer (1988). The German terminology, then, man-

ages to circumvent these theoretical issues by classifying the objects simply as elements.

2.3.3 Compounding

By far the most common forms of word formations in German are compounding and derivation (Donalies calls the latter group implicit derivation). While both can be and often are subsumed under a single category, I follow Donalies here and posit two different categories on the ground that they clearly use a different inventory and because this knowledge is important for learners of German if we hope to provide them with a systematic understanding of German word formation.

Compounds are the result of word formation processes that combine two or more words or confixes.

Explicit derivations are the result of those word formation processes that combine simplex or complex words with derivational affixes.

How to further classify word formation processes is outlined in the following sections.

It has to be noted that in German compounds and derived word forms are written as one word in general. Only in certain circumstances do German writers make use of hyphens to indicate the boundaries of constituents in compounds. This makes it necessary for readers of German texts to develop segmenting skills. Native German speakers are usually able to segment compounds in their individual constituents wihout too much trouble. They have the benefit of having (most of) the simplex forms stored in their mental lexicons and are able to use these, the knowledge on what common word beginnings and endings are, and contextual information to process even complex compounds very quickly. Only occasionally will the processing load be high enough to cause overt manifestations of segmentation problems. This can be the case, e.g., with compounds such

as:

(2.11) Blumentopferde

(2.12) Wehrufer

(2.13) Wachstube

These can cause problems because they are ambiguous with respect to what word boundaries exist. Thus, because *Pferde*(horses) is a valid German word, readers might try to analyse 2.11 as follows:

(2.14) *Blumento - pferd - e*
??? - horse - s
??? horses

Only because no valid segmentations exist for the first part of this compound, readers will be forced to reanalyse the word into three distict constituents:

(2.15) *Blumen - topf - erde*
flower - pot - soil
top soil used in flower pots

It is only possible for readers to know how to segment 2.12 and 2.13 if they know about the context, because each compound can be segmented in two different ways:

(2.16) *Weh - ruf - er*
pain - scream - er
person lamenting

(2.17) *Wehr - ufer*
dam - shore
edge of a dam

(2.18) *Wachs* - *tube*
wax - tube
a tubular container of wax

(2.19) *Wach* - *stube*
watch - room
a guard room

Non-native speakers of German, on the other hand, will find it harder to segment complex compounds correctly. In order to acquire and hone this skill, it is important for learners of German to develop strategies that enable them to quickly analyse words. Because their mental lexicon does not contain all simplex forms of German (at least in the case of beginning and intermediate learners) it is important for them to learn how to recognize common word beginnings and endings. I believe that extensive reading and learning about word formation rules does not only help to improve learners' word segmenting skills, but will also increase their range of receptive and productive vocabulary. I will return to this issue in the following sections.

The following sections provide an overview of the different varieties of compounds that exist in German.

2.3.3.1 Endocentric Compounds

Endocentric compounds are by far the most common forms of compounding in German. They are formed of two parts. In German, in general, the left part modifies the right part. Thus,

(2.20) *Pflaumen* - *baum*
plum - tree
plum tree

(2.21) *Schaukel* - *stuhl*
swing - chair
rocking chair

(2.22) *Auto* - *fahrt*
car - trip
trip with a car

2.20 refers to a particular tree (Baum), one that has plums (Pflaumen) on it; 2.21 refers to a specific chair (Stuhl), one that rocks (schaukeln); and 2.22 refers to a particular kind of trip (Fahrt), one that is undertaken using a car.

Endocentric compounds have some important properties:

- Endocentric compounds are binary. They can always be analysed in two distinct parts, the modifier and the head. Even if wordforms consist of more than two constituents, in case of endocentric compounds, their structure is always binary:

  (2.23) *Apfel* - *kuchen* - *guss*
  apple - cake - icing
  icing for apple cakes

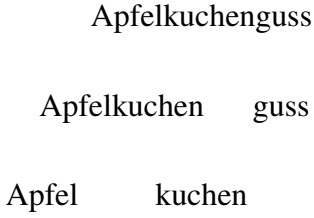


Figure 2.1: Analysis of 'Apfelkuchenguss' - hierarchical structure

The representation with the binary branching tree (figure 2.1) is the more adequate representation of the compound. The compound refers to a certain kind of icing, the icing that is used for a particular kind of cake, a cake that has apples on top of it.

Apfelkuchenguss

Apfel kuchen guss

Figure 2.2: Analysis of 'Apfelkuchenguss' - flat structure

The flat structure (figure 2.2) falsely indicates that the compound has a tripartite structure and that there is no relationship between individual elements on a lower level.

- In German the Right Head Rule applies. This means that the right most constituent of an endocentric compound is the head of the constructions, it determines the nature of the compound as a whole. 2.20 is a tree, 2.21 is a chair, and 2.22 is a trip.

- The preceding constituents are not affected by inflection. It is only the head that undergoes inflectional changes that occur when forming the plural form of nouns, by conjugation, declination, etc. The plural forms of the examples above are: *Apfelbäume, Schaukelstühle, Autofahrten*. Very few exceptions exist to this rule and even these are contentious. For example, even though some speakers of Bavarian German would argue that the plural form of *Semmelknödel* (a dumpling made from left over dinner roles) is *Semmelnknödeln*, one can argue that this plural form goes back to a sketch by the Bavarian comedians Karl Valentin and Liesl Karlstadt that actually takes a tongue-in-cheek look at the rules underlying German plural formation (Valentin, 1978). It would then follow that *Semmelnknödeln* itself can be considered a creative invention that intentionally violates the rules of word formation.

- Attributes of a compound can only serve to provide extra information about the

head. They cannot normally form a relation over the modifier. Using an attribute like *schön* with any of the above compounds, provides additional information about the heads of the compounds, not about their modifiers. *Eine schöne Autofahrt* means a nice trip and does not lead us to conclude about the state or appearance of the vehicle that was used for the trip. Very few real exceptions to this rule exist. While *schnelle Auffassungsgabe* is, indeed, the ability (Gabe) to comprehend something quickly, other constructions, like *rundes Geburtstagskind* are usually formed in order to achieve a surprising effect. *Rund* means round and acts as a modifier of *Geburtstag* (birthday), not *Kind*(child). A "round number" is the number ten and its multiples.

Apart from noun + noun compounds, members of other word classes can function as modifiers and form endocentric compounds with noun heads.

Adjectives:

(2.24) Blaulicht (blue light, flashing on a police car)

(2.25) Magersucht (anorexia, lit.: scrawny addiction)

(2.26) Dunkelmänner (untrustworthy people, lit: dark men)

Verbs:

(2.27) Frisiertisch (vanity, lit.: hair styling table)

(2.28) Brecheisen (crowbar, lit.: break iron)

(2.29) Schwimmring (floatation device, lit.: swim ring)

While the group of noun + noun compounds clearly represents the largest group of endocentric compounds, German speakers/writers also form compounds with adjectival heads, and to a lesser extend compounds with verbal heads. Here are some examples:

(2.30) himmelblau (sky blue)

(2.31) bettelarm (poor as a beggar)

(2.32) mausgrau (grey as a mouse)

(2.33) kontaktschweißen (to contact-weld)

(2.34) spritzgießen (to injection-mould)

Combinations containing verbs that can be shown to have been constructed on the basis of a phrase – it can be argued – are to be considered conversions rather than compounds (Donalies, 2007). Thus, while a phrasal basis for *spritzgießen* cannot be found: (*Sie spritzgießt), a phrase like *Über Winter blieben wir in Mallorca* can be considered a phrasal basis for *überwintern*.

As regards teaching/learning DaF, this seems to be too much nitpicking. For practical reasons, I would argue to treat these cases as endocentric compounds as well.

Other constructions such as *Vergissmeinnicht* (forget-me-not), *Tunichtgut*(good for nothing person), etc., however, overtly show their phrasal nature and will be considered phrases here.

2.3.3.2 Exocentric compounds

Exocentric compounds are a special case of compounding. Contrary to endocentric compounds which are motivated by the fact that they describe the inherent properties of what they refer to, exocentric compounds describe only specific aspects of what they refer to. For example, the following compounds are all endocentric:

(2.35) mausgrau: grey like a mouse

(2.36) Apfelkuchen: a cake made with apples

(2.37) kontaktschweißen: a specific kind of welding

These compounds, on the other hand, are exocentric:

(2.38) Nashorn (rhinoceros, lit.: nose horn)

(2.39) Rotkehlchen (robin, lit.: red throat)

(2.40) Langfinger (thieve, lit.: long finger)

Exocentric compounds form a pars-pro-toto relationship with their referents. The fact that these metonymies are formed after a metaphoric transfer increases the processing load on the side of the hearer/listener, only, of course, if the compound does not exist in her lexicon in unanalysed form. An example will make this clear: while it is obvious that *Schokoladenkuchen* is a kind of cake, hearers/listeners have to be aware of the fact that *Kopf* is frequently used in exocentric compounds to refer to people with a quality that is defined by the modifier.

(2.41) Dummkopf (stupid person)

(2.42) Schlaukopf (intelligent person)

(2.43) Struppelkopf (person with tousled hair)

Without the knowledge that in these cases and words formed analogously the head stands for the person as a whole, it is hard to arrive at a correct interpretation of the compound.

The distinction between exocentric and endocentric compounds, thus, is entirely a semantic one. The group of exocentric compounds is small compared to the endocentric ones. Nevertheless, speakers of Germans form them frequently and learners of German need to know about them in order to be able to interpret ad hoc formations they might encounter. It is also useful to know that the head of exocentric compounds is always a noun.

2.3.3.3 Appositional compounds

Appositional compounds are compounds of two or more words belonging to the same word category which do not have a hierarchical order, or, in other words, they are modifiers and heads at the same time. The majority of appositional compounds consist of adjectives, especially those that refer to colours:

(2.44) rot-weiß

(2.45) schwarz-weiß

(2.46) schwarz-rot-gold

In principle, the order of the elements is arbitrary. Thus, it is possible to call a red and white skirt *rot-weißer Rock* or *weiß-roter Rock*. On the other hand, in many cases it seems that the order of elements is either fixed by convention, or fixed because the order is important. *Schwarz-weiß Fernseher* is the correct way to refer to a black and white TV set and *weiß-schwarz Fernseher* sounds odd or might not be understood at all. *Schwarz-rot-gold* refers to the colours of the German flag, and changing the order of the individual elements would not represent that flag anymore (inverting the order, as a matter of fact, would represent the Belgian flag). While there are examples of verb-verb and noun-noun compounds that appear to have no hierarchical order, these terms, too, seem odd at the least if the order of their constituents element is changed.

(2.47) spritzgießen (injection-moulding)

(2.48) Hausboot (house boat)

While in 2.47 it does not seem possible to say that the process referred to is a specific kind of moulding, but rather that it appears to involve moulding and injecting at the same

time, inverting the elements of the compounds makes the word hard to understand. The same holds for 2.48, which refers to something that is a house and a boat at the same time, rather than a "housy" kind of boat, or the other way around. Still, the existence of one form appears to block other possible permutations, although they would seem inherently well-formed and able to serve as referents for the objects in question. Where order does not seem to matter from a semantic perspective, learners have to find out if there is possibly only one acceptable order that native speakers use. They will most likely be understood if they use a different order and be able to maintain a conversation, but ultimately, to be considered proficient users of German, they will have to use the order that is considered the appropriate one by the language community.

2.3.3.4 Contaminations

Contaminations are words that result from "melting together" two or more words. Contaminations are in most cases ad hoc formations. Individual words are combined into a single form either simply because "it sounds good" or because they have a sequence of sounds in common. It is seldom that the language community starts using contamination which are always formed to achieve a surprising effect. Examples are:

(2.49) Mammufant (Mammut + Elefant)

(2.50) Kurlaub (Kur + Urlaub)

(2.51) philosofaselt (philosophieren + faseln)

Donalies' terminology here is contentious. Elsewhere (Hock & Joseph, 1996), the examples given above would be considered instances of blending together with similar cases in English such as:

(2.52) brunch (breakfast + lunch)

(2.53) motel (motor + hotel)

(2.54) telecast (television + newscast)

Hock & Joseph (1996) reserve the term contamination for words that came into existence because a word co-occurring with another one eventually assumed some of the phonetic properties of its neighbour.

(2.55) Protoromance: grevis (Latin: gravis and levis)

(2.56) English: female (French: male and femelle)

These changes are usually due to a reanalysis that occurs when the terms are introduced as loan words into another language.

2.3.3.5 Reduplications

Reduplication is a form of compounding where a word or morpheme is combined with itself. A vowel or the stem is usually altered in the second part of the resulting compound. Reduplication is not a very productive word formation process. The resulting compounds are all restricted to the informal register. Examples include:

(2.57) Schickimicki (fancy)

(2.58) Krimskrams (bric-a-brac)

(2.59) Wirrwarr (chaos)

Donalies (2007) does not consider what she calls echo words forms of reduplication and analyses them as endocentric compounds. Thus *graugrau* is a specific kind of grey, a very grey grey, *Film-Film*, a term coined by the German TV network Sat1, presumably refers to a special weekly feature film: thus a very 'filmy' kind of film.

2.3.4 Explicit derivations

Explicit derivations, I would argue, is the class of word formation processes that are ideally suited for systematic learning. As can be seen from the table at the end of this section, the number of affixes that German uses in derivational processes is fairly limited. Each of these affixes can occur only with a restricted set of bases and will for the most part have a predictable effect. Affixes used in derivation, although their semantics might be considered more abstract than that of concrete nouns for example, still have meaning. Once learners know what bases a certain derivational affix can attach to and what meaning it contributes to the meaning of the word form as a whole, they are able to analyse other words containing the same affix, infer the meaning (if they know the meaning of the base) or predict it, since they already know part of it. Moreover, they can also use these affixes creatively to form novel word forms.

Derivation in German is used to derive mostly nouns, verbs and adjectives. The bases that derivational affixed attach to are also for the most part nouns, verbs and adjectives.

Nouns can be derived by attaching prefixes, suffixes and circumfixes to a base. The prefixes that can attach to a noun base can be further classified in falsificative and augmentative prefixes. The former class is used to contribute the meaning of absence to a word form. By attaching it to a base, it is indicated that a property, quality, etc. denoted by the base is missing, or that the wordform refers to just the opposite of what the base refers to:

(2.60) Unwort

(2.61) Antikommunist

(2.62) asozial

2.60 refers to a non-word, it is used to refer to words that the speaker/writer considers

abominations of the German language that should not exist at all. Every year a list of *Unworte* of the year is assembled, usually containing words coined by politicians or in administrative contexts that are felt to be insulting or dehumanizing, e.g., *Babyklappe* (baby hatch) to refer to a part of a hospital where people in need can anonymously drop off a baby.

2.61 refers to someone who is against communism and 2.62 refers to behaviour that is violating social norms. As can be seen from these examples, the degree of how easy it will be for a learner to establish the correct meaning of a certain derivation varies. In some cases it is necessary to know the cultural context of a word, in others, knowing about the meaning of the individual parts suffices to infer the meaning of the form as a whole.

Augmentative prefixes do not invert the meaning of the base, rather they add another semantic feature to the feature set of the base.

(2.63) Megaparty

(2.64) Erzhalunke

(2.65) Vizeweltmeister

In 2.63 and 2.64 the prefixes both contribute the meaning of intensity to the forms as a whole. A *Megaparty* is perceived to be a very big and exiting party. An *Erzahalunke* is still a villain, but even more so than just a 'normal' *Halunke*. *Vize-* on the other hand, adds the meaning of "secondness" to the form as a whole. *Weltmeister* is the world champion, but 2.65 is second only to it.

Suffixes used to derive nouns fall in a number of different subclasses:

Nomina agentis: The most productive suffix deriving nouns is *-er* that can attach to just about any verb to denote the person, or object that carries out the action described by the verbal base.

(2.66) Helfer (helper)

(2.67) Geber (giver)

(2.68) Flammenwerfer (flame thrower)

An alternative form is the foreign loan suffix -ant that mainly attaches to Latinate verb bases.

(2.69) Gratulant (someone congratulating)

(2.70) Denunziant (someone denouncing)

(2.71) Konfirmant (someone taking part in confirmation, the religious ceremony)

It is a matter of choice whether the nouns derived by attaching *-er* to noun bases should also be called nomina agentis, since the bases do not refer to an action, the derived noun as a whole, on the other hand, refers to an action that is directed at or at leas somehow connected to the base.

(2.72) Gärtner (Garten + er, gardener)

(2.73) Metaller (member of the metal workers' union)

(2.74) Rentner (Rente + er, pensioner

Nomina patientis is the name for nouns that refer to persons or objects undergoing the action described by the base.

(2.75) Konfirmand (someone undergoing the rite of confirmation)

(2.76) Lutscher (sucker, candy)

(2.77) Schützling (Schutz + ling, protégé)

Expressive nouns are formed by adding the suffix *-er* to a verb in order to form a noun that refers to an utterance. This is a relatively small class of words.

(2.78) Rülpser (rülpsen + er, a burp)

(2.79) Schmatzer (schmatzen + er, smack of one's lips, also: a kiss)

Motiva are formed by attaching a suffix to a base that is not marked for gender in order to mark it for a certain gender.

(2.80) Frisörin (also: Friseuse, Frisör + in, female hairdresser)

(2.81) Abteilungsleiterin (Abteilungsleiter + in, female department head)

(2.82) Enterich (also Erpel, Ente + rich, male duck)

Of course, all German nouns are marked for gender, but many nouns that are grammatically marked as masculine, especially the ones referring to jobs, are gender neutral in the sense that they can be used to refer to both male and female members of a profession. Although there has been a tendency to use motiva in the last few decades, some nouns seem to be exempted from this word formation process. For example, an expert in a certain field is often referred to as *Fachmann* in professional contexts, although *Fachfrau* is a valid alternative:

(2.83) Hotelfachmann (less common: Hotelfachfrau)

(2.84) Reiseverkehrsfachmann (less common: Reiseverkehrsfachfrau)

(2.85) IT-Fachmann (less common: IT-Fachfrau)

In other cases, completely new words were coined, because the traditional word presumably cannot be interpreted as being gender-neutral since one of the constituents refers to a female, for the most part.

(2.86) Geburtshelfer (Hebamme, mid-wife)

(2.87) Krankenpfleger (Krankenschwester, nurse)

(2.88) Raumpfleger (Putzfrau, cleaning lady)

Diminutives are formed by adding a suffix to a noun base. The affix adds the semantic feature of smallness or quaintness to the form as a whole. Suffixes used to form the diminutive forms can vary regionally. In the south of Germany, for example, the suffix *-le* is frequently used, whereas in Hessia, speakers frequently use *-che(n)* in diminutive forms. The standard variation, for the most part, uses *-chen* and *-lein*.

(2.89) Bäumle, Bäumche, Bäumlein (little tree)

(2.90) Häuschen, Häuslein (little house)

(2.91) Bierchen (small beer, or rather only one beer instead of a large number)

There are also a few lexicalised forms that are based on diminutive forms that have assumed an idiosyncratic meaning.

(2.92) Mädchen goes back to Maid (maiden) + chen. The term is now used to refer to any young female (girl)

(2.93) Fräulein (Frau + lein) does not refer to a little woman, but rather to an unmarried woman. It is considered old-fashioned by most speakers of German

Nomina qualitatis are nouns that refer to a quality denoted by the adjectival base. In German the suffixes *-heit, -keit* and *-nis* are the most productive suffixes used to form nomina qualitatis.

(2.94) Unverschämtheit (outrageousness)

(2.95) Sauberkeit (cleanliness)

(2.96) Wildnis (wilderness)

The only circumfix that is used in German in the derivation of new nouns is *ge-...-e*. It can be attached productively to most verb bases in order to form nouns referring to the action denoted by the verb.

(2.97) Geheule (ge + heulen + e, whining, lamenting)

(2.98) Aufgereiße (ge + auf + reißen + e, picking up (people in a bar, for example)

(2.99) Geschreibe (ge + schreiben + e, writing, scribbling)

All of these nouns have a negative connotation. Note also that the circumfix separates the prefix from verbs with separable prefix. The same circumfix can also be used to form collective terms togther with noun bases. This process is only mildly productive.

(2.100) Gebirge (ge + Berg + e, mountain range)

(2.101) Gestänge (ge + Stange + e, an assembly of beams)

(2.102) Gebälk(e) (ge + Balken + e, the beams under the roof of a building)

Like nouns, most adjectives can be used together with a falsificative or augmentative prefix.

(2.103) ungeil (un + geil, not cool)

(2.104) apolitisch (a + politisch, apolitical)

(2.105) hypergenial (hyper + genial, extraordinary)

The suffixes that can be used to derive adjectives can be classified according to their semantic contributions to the resulting wordforms:

potential: Suffixes that add the meaning of potential to the base include *bar, sam* and *-abel*.

(2.106) trinkbar (drinkable)

(2.107) lehrsam (displaying willingness to learn)

(2.108) indiskutabel (out of the question)

Adding *-bar* to *trink-* renders a wordform that indicates the potential of something to be drunk, *-sam* together with the base *lehr-* indicates the potential to be taught, and *-abel* added to *diskut-* would refer to something that has the potential of being discussed. This word, though, does not exist in German, a second derivation – adding the prefix *in-* – is necessary to arrive at a word that is found in standard German, meaning lacking the potential of being discussed.

identity, partiality, or resemblance The suffixes *-ig*, *-isch* and others can be added to noun bases in order to indicate that the word attributed by the adjective equals, partly consists of or resembles what the noun base of the adjective refers to.

(2.109) holzig (is like or partially consists of wood)

(2.110) euphorisch (to be in an euphoric state)

(2.111) nebulös (appears to be enveloped in fog)

diminutive: Adding the suffix *-lich* to an adjectival base indicates that the quality resembles that of the adjective to some extend.

(2.112) gelblich (yellow-ish, somewhat yellow)

(2.113) weißlich (white-ish, somewhat white)

(2.114) süßlich (somewhat sweet)

negation: The suffix *-los* can be added to a wide variety of nouns to indicate the absence of what is denoted by the noun base.

(2.115) schuldlos (without guilt)

(2.116) arglos (without expecting harm)

(2.117) führerlos (without leadership)

transposition: Donalies (2007) argues that in some cases the suffix has no semantic contributions to make to the resulting adjective. It's sole purpose in these cases is to transpose the base, so it can be used in the syntactic context that requires an adjective. Examples of transositions are:

(2.118) folgsam (folgen + -sam, the qualitiy of obeying)

(2.119) sterblich (sterben + lich, mortal, having the property of eventually having to die)

(2.120) hinderlich (hindern + lich, hindering, having the quality of hindering)

It should become clear that the way word formations are assigned to specific categories can appear arbitrary at times. The fact remains, however, that knowing about the semantic properties of derivational affixes, however abstract this might be, is essential to infer the meaning of ad-hoc formations or of complex forms that we have not encountered before.

The only circumfix used in adjectives is *ge-. . . -ig*. This does not appear to be productive anymore, however.

(2.121) gefügig (compliant)

(2.122) gehässig(spiteful)

(2.123) gelehrig(able to learn)

Constructions such as *gehörnt* (equipped with horns) and *behaart* (equipped with hair) are sometimes considered pseudo participles and analysed as derivations using circumfixes. I follow Donalies (2007) who considers them adjectival conversions of verbs on the grounds that the underlying verbs are possible words:

(2.124) *Die Ehefrau hörnte ihren Mann.*
The wife horned her husband.
The wife committed adultery.

(2.125) *Der Geigenspieler behaarte seinen Bogen.*
The violin player haired his bow.
The violin player equipped his violin bow with horse hair.

It is the verbs that have the richest inventory of prefixes. These again can be grouped into the following classes:

negations: The prefixes *miss-* and *ver-* attach to verb bases and serve to negate the meaning of this base.

(2.126) missdeuten (misinterpret)

(2.127) misstrauen (mistrust)

(2.128) verspielen (gamble away)

ornative: Prefixes attaching to adjectival bases indicate that the action denoted by the verb serves to add the quality denoted by the base.

(2.129) befreien (be + frei + en, to free, make free)

(2.130) verhärten (ver + hart + en, to make hard)

(2.131) ermutigen (er + mutig + en, to encourage)

(the *-en* in all of these examples is not a derivational suffix, it is the infinitival inflection)

privative: Prefixes of this class serve just the opposite function than ornative prefixes. They indicate that the verb serves to take something away, or diminish in terms of a certain quality.

(2.132) entkleiden (undress)

(2.133) entkorken (de-cork)

(2.134) entführen (kidnap, hijack, lit.: to lead away)

instrumental: These prefixes serve to form verbs that indicate what instrument is used to perform an action.

(2.135) erdolchen (to stab, to knife)

(2.136) verschrauben (to fasten with screws)

(2.137) bepinseln (to colour with a paint brush)

comparative: Prefixes in this group serve to form verbs that indicate that the action denoted resembles the qualities of the noun base of the word form.

(2.138) bemuttern (to care for like a mother)

(2.139) verbrüdern (to unite, like brothers)

(2.140) ermannen (to summon one's courage, like a (real) man)

factive: Verbs belonging to this class indicate a change of state. The base in these derivations indicates the end stage, while the verb as a whole serves to denote the process that leads up to this final stage.

(2.141) verwüsten (destroy, lit.: to turn into a desert)

(2.142) verblöden (to become dumb, turn into an idiot)

(2.143) verfilmen (to turn into a movie)

deictic: A group of prefixes can attach to verbs and contribute a deictic meaning to the resulting word form. While the base denotes an action that does not necessarily indicate a certain direction or destination, the resulting wordform is more restricted in terms of temporal or spacial terms.

(2.144) beschauen (be + schauen, inspect, watch intensely)

(2.145) bereisen (travel to a certain place)

(2.146) beleuchten (shine the light on something)

inchoative: These prefixes attach to verb bases and contribute the meaning of beginning, setting in.

(2.147) erklingen (to start to ring)

(2.148) erschallen(to resound)

(2.149) erblühen(to start to bloom)

The number of suffixes used in the derivation of verbs is fairly small. For the most part, they are used to attach to confixes and other loan element in order to use them as verbs.

(2.150) infizieren

(2.151) intensivieren

(2.152) sondieren

It is also possible to form diminutive forms of some verbs by adding *-ln* to a verbal base.

(2.153) tröpfeln (tropfen + ln, to drizzle)

(2.154) lächeln (lachen(laugh) + ln, to smile)

(2.155) spötteln (spotten(ridicule) + ln, to mock)

Most of the wordforms in this category can be considered established to the point of lexicalization, i.e., the words seem to be stored as single units in a native speaker's lexicon and have taken on an idiosyncratic meaning, although this might still be related to the meaning of the base.

There are also two circumfixes that can be used to derive verbs from adjectives and nouns in German. *be-...-ig* can be used with nouns, in which case it has an ornative character, or it can be used with adjectives, in which case it is factive.

(2.156) beleidigen (be + Leid + igen, to give sorrow to, to insult)

(2.157) beerdigen (be + Erd + igen, to put in the ground, to bury)

(2.158) beschleunigen (be + schleunig +en, to make fast, accelerate)

The circumfix *in-...-ier* and its allomorph *in-...-isier* derive a verb that denotes a specific location that is involved in the action.

(2.159) inthronisieren (enthrone)

(2.160) inszenieren (to put into scene, to stage)

Finally, there is a group of suffixes that can be used to derive adverbs from nouns and adjectives. Here are some examples for the suffixes *-wärts, -halber*, and *-ns* which attaches to the superlative form of some adjectives.

(2.161) himmelwärts (toward the sky)

(2.162) umstandshalber (because of the current circumstances)

(2.163) schnellstens (in the fasted manner possible)

Table 2.1: German affixes

Affix \Base	**Noun**	**Adjective**	**Verb**	**Confix**	**Other**
a-	Amoral	ahistorisch			
-abel	profitabel		spendabel	diskutabel	
-ade	Robinsonade			Marinade	
-age	Spionage			Montage	
-aille				Journaille	
-al	Personal			Signal	
-al	horizontal			global	
-alie	Archivalie		Freassalie	Mineralie	
an	Analphabet	anorganisch			
-and	Doktorand			Konfirmand	
-ant	Asylant		Bummelant	Intrigant	
-anz				Intriganz	
-ar	Bibliothekar			Kommentar	
-är	Funktionär			Sekretär	

Continued on next page

Affix \Base	Noun	Adjective	Verb	Confix	Other
-ast				Phantast	
-at	Rektorat			Stipendiat	
-ation				Integration	
-bar	fruchtbar	offenbar	verrückbar		
be-	befreunden	befreien	besuchen		
be-...-ig	beaufsichtigen	begradigen			
-chen	Kindchen	Sensibelchen			
de-	Demission	dezentral	demaskieren		
des-	Desinteresse	desorientiert	desorganisieren		
dis-	Disharmonie	disharmonisch	disorientieren	dissoziieren	
-e		Süße	Leuchte	Geologe	in Bälde
-ei	Gärtnerei		Blödelei		
-el	Bändel				
-el	frösteln	fremdeln	lächeln		
-elei	Diebelei		Liebelei		
-ell	sensationell			individuell	

Continued on next page

Affix \Base	Noun	Adjective	Verb	Confix	Other
-en	golden				
-ens			vergebens		
-ent	entgräten	entblöden	entladen		
-ent				intelligent	
-enz				Intelligenz	
-er	erkunden	erbittern	erdenken		
-erei	Käserei		Spotterei		
-erich	Enterich	Dummerich	Flatterich		
-erie	Szenerie			Hysterie	
-ern	gläsern				
erz-	Erzrivale	erzböse			
-esk	clownesk				
-ess	Stewardess				
-esse		Akuratesse			
-eur	Pamphleteur	Bankrotteur		Friseur	
-euse	Balleteuse	Bankrotteuse		Friseuse	

Continued on next page

Affix \Base	Noun	Adjective	Verb	Confix	Other
ge-...-e	Gerede				
ge-...-ig			gefügig		
-haft	pinguinhaft	krankhaft	schwatzhaft		
-heit	Menschheit	Zartheit			
-i	Hirni	Dummi	Brummi	Prolli	
-iade	Schubertiade				
-ian		Blödian	Schlendrian		
-ibel				disponibel	
-ice				Direktrice	
-icht	Röhricht	Dickicht	Kehricht		
-ie	Aristokratie	Anomalie		Ironie	
-ier	Kanonier	Privatier			
-ier	gastieren	halbieren		diskutieren	
-ifizier	personifizieren	diversifizieren		identifizieren	
-ig	ängstigen	reinigen			
-ig	eisig	völlig	wendig		sofortig

Continued on next page

Affix \Base	Noun	Adjective	Verb	Confix	Other
-igkeit		Frömmigkeit			
-ik	Methodik			Thermik	
in-		indiskutabel			
-in	Mörderin				
-ine	Dackeline	Blondine			
-inski		Radikalinski			
-ing			Stretching		
inter	Intertext	intersprachlich			
-ion	Institution	Abstraktion		Kreation	
-isch	launisch	genialisch	misstrauisch	elektrisch	
-isier	rivalisieren	privatisieren		sympathisieren	
-ismus	Terrorismus	Rationalismus	Abspaltismus	Chauvinismus	
-ist	Saxophonist	Purist			
-it				Kosmopolit	
-ität	Moralität	Rarität		Authentizität	
-itis	Telefonitis	Banalitis		Arthritis	

Continued on next page

Affix \Base	**Noun**	**Adjective**	**Verb**	**Confix**	**Other**
-iv	effektiv		relativ		
-keit		Lächerlichkeit			
ko-	Koautor	koevultiv	koexistieren		
-lein	Kindlein				
-ler	Postler		Abweichler		
-lich	freundlich	bläulich	bedrohlich		widerlich
-ling	Dichterling	Naivling	Lehrling		
-lings	bäuchlings	blindlings			
-los	treulos		reglos		
-ment				Engagement	
miss-	Missgeburt	missvergnügt	misstrauen		
-ner			Redner		
-nis	Bildnis	Finsternis	Erlaubnis		
-o	Krawallo	Realo		Prolo	
-oid	Planetoid				
-oid	grippoid	technoid		paranoid	

Continued on next page

Affix \Base	Noun	Adjective	Verb	Confix	Other
-or				Investor	
-ös	skandalös				
-ose	Tuberkulose				
para-	Paraästhesie	paramilitärisch			
post-	Postkubismus	postpubertär			
prä-	Präpubertät	präpubertär	prädisponieren		
re-	Reanalyse	reaktiv	reokkupieren		
-s	Knicks				
-s		bereits			
-sal	Mühsal	Trübsal			
-sam	tugendsam	langsam	folgsam		
-schaft	Lehrerschaft	Bereitschaft	Belegschaft		
-sel			Überbleibsel		
-ski		Besoffski			
-t			Fahrt		
trans-	Transsibirien	transatlantisch			

Continued on next page

Affix \Base	**Noun**	**Adjective**	**Verb**	**Confix**	**Other**
-tum	Heidentum	Heiligtum	Irrtum		
un-	Untiefe	unschön			
-ung	Stallung	Festung	Bedeutung		
ur-	Urwald	uralt			
-ur	Agentur			Frisur	
ver-	vergolden	versüßen	versuchen		
zer-	zerbeilen	zerkleinern	zerpflücken		

Note. Adapted from Donalies (2007, pp. 133-137)

2.3.5 Conversion

Conversion is the term used to describe the transposition of a word belonging to some word class to another one, a process that is not overtly marked on the morphological level. In German, all parts of speech can be converted into a noun.

(2.164) rechnen, das Rechnen (the calculating)

(2.165) hier, das Hier und Jetzt (the here and now)

(2.166) das, das Das am Satzanfang (the the at the beginning of the sentence)

In addition, participles can be used as adjectives:

(2.167) Es war ein gelungener Spielzug. (It was a successful move.)

(2.168) Der lachende Sieger (the laughing winner)

(2.169) Das gebrochene Versprechen (the broken promise)

All adjectives at all levels of comparison can be converted into nouns, some adverbials can be used as nouns, too, and so can entire phrases.

(2.170) Das Blau des Himmels (the blueness of the sky)

(2.171) Das Allerschönste im Leben (the best thing in life)

(2.172) Das Nebeneinander von Arbeit und Privatleben

(2.173) Das Sich-auf-die-eigene-Schulter-klopfen ging uns auf die Nerven.

Donalies (2007) points out that even affixes can be used in conversions:

(2.174) Sozialismus, Kommunismus und andere Ismen (from *-ismus*)

2.3.6 Implicit derivation

Implicit dervation is a word formation process that can no longer be considered productive. It refers to the formation of causative verbs that involves a stem vowel change of the base. The most important implicit derivations are:

(2.175) tränken (soak, from *trinken*, to drink)

(2.176) senken(to lower, from *sinken*, to sink)

(2.177) setzen(to put, from *sitzen*, to sit)

(2.178) legen(to lay, from *liegen*, to lie)

2.3.7 Reductions

Reductions include word forms that are the result of shortening a long form:

(2.179) Schupo (from Schutzpolizei, a branch of the police)

(2.180) Mathe (from Mathematik)

(2.181) Lise (from Elisabeth)

Acronyms are also reductions. They refer to abbreviations that are actually pronounced as such. There are, in principle, two types of acronyms. Some that are pronounced as words:

(2.182) RAF (Rote Armee Fraktion, German terrorist organisation)

(2.183) GAU (Größter anzunehmender Unfall, worst case scenario)

(2.184) FAZ (Frankfurter Allgemeine Zeitung, newspaper)

and also those that are pronounced letter by letter:

(2.185) LKW (**L**ast**k**raft**w**agen, truck)

(2.186) SPD (Sozialdemokratische Partei Deutschlands, political party)

(2.187) ARD (Allgemeine Rundfunkanstalten Deutschlands, broadcasting network)

Abbreviations that are only used in written German and are not pronounced as such in spoken language are not considered reductions:

(2.188) Dr. (doctor)

(2.189) Hr. (Herr)

(2.190) Frl. (Fräulein)

2.3.8 Remotivations and play-on-words

The final word formation processes that are discussed here are remotivations and play-on-words.

Remotivations are word forms that are formed by analogy because language users either by accident or deliberately misinterpret the origins of the original word and adapt it to a specific context. It is easiest to illustrate this with some examples.

(2.191) Maulwurf (mole) can be traced back to Middle High German *moltwerfe* (someone throwing dirt). It has been reanalyzed as follows:

(2.192) *Maul* - *wurf*
muzzle - throw
someone throwing with their muzzle

(2.193) Hängematte (hammock) was formed on the basis of *hamaca* and literally means hanging mat. The phonetic form of the original word, thus lead to a German reinterpretation

(2.194) Pizza Hut (pronounced: hoot) was the word many Germans first used when the American fast food chain opened its first restaurants in Germany. They associated the form of the red brick roof with a hat (German: Hut)

(2.195) Tollpatsch (a clumpsy person) used to be written like the Hungarian word it originates from: Tolpatsch. The Hungarian word referred to a type of soldier. German speakers associated the first part of the word with the German adjective *toll* (crazy). The reformers of German orthography acknowledged this fact by ruling that the form with the double l ought to be the only officially correct one (confer *Duden. Die deutsche Rechtschreibung*, 1996).

Play-on-words, like remotivations, do not follow a set of rules, neither are they predictable like derivations. Since they are creative formations of new words, however, they can be considered to fall into the domain of word formations. Examples for play-on-words are:

(2.196) Ernstbold (serious person, formed in analogy to Witzbold, a funny person)

(2.197) Obertan (ruler, formed in analogy to Untertan, subject)

(2.198) Schwarzhören (using a radio without paying public broadcasting fees,formed in analogy to schwarzfahren, using public transportation without paying)

The last example refers to a term coined by the GEZ, the organisation that is in charge of collecting the fees for the public broadcasting networks from all households that own a radio and/or TV set.

2.3.9 Summary

German word formation is rule governed. Although there are various different processes that have been or are still productive, as far as derivation goes, they are based on a fairly limited set of derivational affixes. While some word formations in German, as in any other natural language are no longer transparent, or may not be easy to interpret without contextual and/or cultural information, most word formations are analysable in the sense that the learner is able to infer the meaning of a complex form from the semantics of its constituents or do so at least to some extend. The language community is creative. Word formation is one of the most important areas where the creativity of language users shows. It is up to the language community whether a novel word form will become part of the vernacular or not be picked up at all. Participants in a conversation will always face the task of analysing unknown word forms, no matter if they are ad-hoc formations or just forms that have not been encountered before. For learners of German, to succeed in their language acquisition, it is important that they become acquainted with at least the productive and semi-productive word formation processes, learn about the meanings of derivational affixes, as well as confixes, and to learn not only to analyse complex wordforms, but also to productively use the system of word formation to achieve communicative goals in situations of authentic language use.

Having covered the linguistic foundations that are important for the dissertation, the next section will be concerned with the second language acquisition process. While the acquisition of word formation rules will be of some concern here, the emphasis will be placed mainly on the acquisition of vocabulary in a second language.

2.4 Vocabulary acquisition in a foreign language

How vocabulary is learned, or acquired is studied by vocabulary acquisition researchers. Their research, methodologies and insights will be the subject of this part of the dissertation. As the main concern is with German as a foreign language, I will concentrate on the acquisition of vocabulary in a foreign language. This area itself is a subdiscipline of second language acquisition research. The section will start out with an overview of the latter and will then discuss some influential theories and research results that are pertinent to the project.

2.4.1 What is SLA?

> The study of S[econd-]L[anguage] A[cquisition] is a broad, interdisciplinary field of inquiry which aims to describe and explain the development and non-development of languages and language varieties beyond the first language. SLA researchers study children and adults learning naturalistically or with the aid of formal instruction, as individuals or in groups, and in foreign, second-language, and lingua franca settings. The research draws upon and contributes to knowledge and procedures in a variety of disciplines, including theoretical linguistics, neurolinguistics, psycholinguistics, sociolinguistics, historical linguistics, pidgin/creole studies, applied linguistics, psychology, sociology, anthropology, and education. SLA research findings are used to test hypotheses and build theories in those areas, as well as for a variety of practical purposes such as the improvement of language teaching, language testing, teacher education, and the design of instructional programs delivered through the medium of a second language or dialect.
>
> Source: Larsen-Freeman & Long (1990)

The study of second language acquisition(SLA) is concerned broadly with the processes involved in the acquisition of a language by an individual other than the first language and the contexts in which this acquisition takes place.

The term second language acquisition itself is slightly misleading as it encompasses not only the acquisition of the first language learned after the mother tongue has been acquired, but also the acquisition of subsequent languages. Clearly coined in a monolingual context, moreover, it does not account for the fact that monolingualism on a global scale is an exception rather than the norm. In many parts of the world people are exposed to and expected to use more than one language from early childhood on, so it would be difficult to answer what their native tongue is.

SLA looks at the development of learners' ability to use a second language. This ability is called linguistic competence (Chomsky, 1965) by researchers in the Chomskian tradition. Communicative competence is the term those researchers prefer to use who believe that the former term is too narrow as it excludes the social function of language and the importance of society and the learner's role in the learning environment (Hymes, 1992).

In the literature, frequently, a distinction is made between language learning and language acquisition, usually taking the former as a process that involves learning of rules and specifically focusing on individual language phenomena while the latter is usually used to refer to processes which do not necessarily require overt attention on specific phenomena. I will follow R. Ellis (2008) here and use both terms interchangeably. He writes:

> A distinction is sometimes made between 'ACQUISITION' and 'LEARNING' (for example, Krashen, 1981). The former refers to the subconscious process of 'picking up' a language through exposure and the latter to the

> conscious process of studying it. According to this view, it is possible for learners to 'acquire' or to 'learn' L2 features independently and at separate times. Although such a distinction can have strong face validity – particularly for teachers – it is problematic, not least because of the difficulty of demonstrating whether the processes involved are or are not conscious. (R. Ellis, 2008, p. 7)

The distinction implicit vs. explicit learning is a useful one. By and large, implicit learning is defined, according to R. Ellis (2008), as taking place without awareness or intentionality. He cites studies showing that learners were shown to have acquired new vocabulary after completing a task that required them to read a text for meaning, but without being warned that they would be tested on new vocabulary. Explicit learning, on the other hand means concentrating on specific language phenomena, and therefore involves intentionality. Awareness, in his view, is problematic as there is no way of controlling for participants' awareness in experimental settings.

A similar pair of terms is intentional vs. incidental learning. According to R. Ellis (2008) it is used by fewer researchers and in less studies than explicit and implicit. It appears that the distinction between both pairs lies foremost in the design of experiments and the role of awareness. In experiments that look for evidence of explicit learning, the participants are given explicit rules or they are asked to deduce them on the basis of data they are provided with. In the case of intentional learning, participants are asked to memorize new language items. Neither the instruction nor the task has a focus on rules. While tasks in experiments concerned with incidental learning ask participants to concentrate on a specific language phenomenon and later testing another one, experiments in the area of implicit learning do not try to "deceive" participants, but ask them to try to process all the input they are getting.

Awareness seems to be the focus in the implicit/explicit paradigm and the experiments are designed to control for it. Schmidt (1990) calls into question the notion that learning without awareness can take place at all. His model (Schmidt, 1994, 2001) postulates awareness as noticing and metalinguisitic awareness. Noticing involves the attention to what he calls "surface elements." Metalinguistic awareness is the awareness of the processes that are involved in incorporating new information into memory. R. Ellis (2008) consequently suggests to redefine implicit learning as learning that takes place without metalinguistic awareness.

The debate over the importance of awareness for language learning is far from over. Schmidt (2001) writes:

> Both implicit and explicit learning surely exist and they probably interact [...] What these two kinds of learning, implicit and explicit, have to do with each other continues to be a topic of great debate within SLA and elsewhere. In SLA the question has frequently been posed in terms of whether or not 'learned' knowledge can become 'acquired' or whether the learner's conscious hypotheses can become internalized [...] Another, possibly more productive, way to pose the question is in terms of learning processes (rather than types of knowledge), to ask whether bottom-up, data driven processing, and top-down, conceptually driven processing guided by goals and expectation (including beliefs and expectations concerning the target language grammar), interact; to which the answer is probably yes, they do.

While awareness is a concept that is laden with philosophical issues, Schmidt continues to examine the role of attention which he considers more fruitful. In his view, it is vital that learners attend to input (aware or unaware). Noticing, using his definition, is the first step of processing. Learners do not notice "raw input" but certain elements in it.

This initial processing is what many researchers (Schmidt, 2001; Gass & Selinker, 2008) consider vital for the acquisition of a new form.

Awareness still continues to be an area of SLA inquiry. Hama & Leow (2010) report on a study that replicates the study by Williams (2005) which, as the authors admit, received some criticism. While Williams had to show that learning without awareness is possible, critics questioned that the results of his experiment validated this conclusion. Hama & Leow (2010) repeat the study with some minor changes. Subjects are taught a set of artificial determiners. They are informed on their semantic meaning, but the information that each semantic meaning is represented by a different word, depending on whether it occurs together with an animate or inanimate object is withheld. Another feature that the participants are not informed about is the encoding of relative proximity or distance of the object the determiner is used with. With a number of methods to distract participants from these facts, learners are first presented a set of "correct" sentences using the new determiners and are later asked to guess which determiners would best be used in gap sentences. It is hard to imagine, though, given that the only type of input that has a very high saliency in the first part of the experiment are the artificial determiners – the rest of the sentence consisted of regular English words – that the participants could be tricked into being unaware of the novel forms.

Personally, I concur with Schmidt. Studying the role of attention for the learning process appears to be more interesting as it can potentially yield important insights into how to influence learners attention in ways to benefit their learning process. Attempting to show that awareness is not necessary for learning is only of academic interest and the results of studies existing in this area, seem to be of a questionable nature.

Although R. Ellis (2008) concludes that no studies are currently available on the effects of implicit vs. explicit learning, the majority of studies he evaluates indicate that the learning outcome of explicit learning are higher or at least equal to implicit learning.

None of the studies indicated that implicit learning was superior to explicit learning, at least short term.

This result might appear unspectacular. Rote learning of vocabulary has been an integral part of the majority of language courses and learners have done reasonably well using intentional learning (in the sense that learning a certain list of vocabulary is the specific learning goal). Why is this dissertation then concerned with extensive reading and incidental vocabulary acquisition (in the sense that new vocabulary might be learnt as a side effect of another activity)? There are at least two good reasons, and the remainder of this chapter is intended to show their validity. On the one hand, there is motivation: in the section on motivation, I will argue that motivation is connected to goals. Concrete goals, in general, lead to a higher motivation than abstract goals. Learning 150 new words in a week, although it might seem like a realistic goal, is abstract. The result of learning a list of 150 words is immaterial for the most part. A goal along the lines of being able to read a newspaper article, or being able to order a meal in a restaurant, is more concrete. While the latter of these two goals might be achieved by rote learning a few phrases, the former is a long term goal. Learners with a long term goal will invest a considerable amount of time into achieving it. Their motivation can be maintained by making their progress transparent. Concrete outcomes, here, are again more important than abstract ones. In general, the motivation will be higher if learners realize that they are able to read a timetable, understand the weather forecast or understand a joke. Reading in a foreign language is both a way of learning, and a way of maintaining learners' motivation.

As for the second reason: in the section of vocabulary acquisition it will become clear that knowing what a certain word means is a complex concept. For example, words usually have different meanings in different contexts, occur in certain collocations, have various degrees of acceptability depending on register, text type and social setting. To come to a full understanding of what a word means, or can potentially mean, learners

need to be exposed to it in as many different contexts as possible. Reading is the most efficient way to achieve this. Learning with a focus on a certain goal (intentional learning) is necessary, and QuickAssist enables learners to access concrete information on the meaning, the distribution, and other features of particular words. But only when intentional learning is accompanied by exposing learners to large amounts of authentic input can they increase their proficiency in a language. And while some facts about a word will be learned intentionally (by memorizing a dictionary definition, for example), many of the finer nuances will only be learned incidentally, because they are not covered in a language class, a textbook or the dictionary that the learner consults.

Although the terms have their own inherent problems, SLA often distinguishes second language learning and foreign language learning. Here, the terminology defines the context in which learning is taking place. While the former describes a learning situation in which learning takes place in an environment in which the language to be learned is used by the majority of people to communicate on an everyday basis, the latter term is used to refer to a learning environment in which the language to be learned is not the language people use in general to communicate. While these terms can be applied to a British au-pair learning German in Austria (second language learning) or a Chinese student taking Croatian lessons at a Chinese university (foreign language learning), problems will arise for example when talking about learning processes taking place in multilingual environments, e.g., learning German in Waterloo with a sizeable German speaking minority.

Another dichotomy that is often used in SLA is naturalistic versus instructional acquisition. It might be the case that the majority of learners either learn a second language without formal instruction in a context where the second language is used in their environment to communicate, e.g. immigrant workers, or in their native country by attending language classes. However, with the vast majority of SLA studies carried out at universities, the participants of theses studies can be safely assumed to be university students who

are unlikely to have only a naturalistic or instructional language learning history.

To re-cap some of the important aspects that have been discussed so far:

SLA has attempted to answer questions such as the following:

- Does the acquisition of a second language differ from the acquisition of the first language, and if so: how do these processes differ?
- How can linguistic/communicative competence be operationalized and measured?
- Is there a specific order in which learners develop this competence?
- How close can a learner's competence get to that of a native speaker (ultimate attainability)?
- How does age, instruction, intelligence, the native language, previous language learning experience, etc. influence the learning process?
- How do insights about second language acquisition contribute to our understanding of the brain/mind?
- What impact do SLA research findings have on foreign language didactics?

Some SLA terminology was introduced. This terminology is used by researchers to define the purpose of a study, a hypothesis, etc. because they offer abstractions away from mental processes, specific learning settings etc. in order to capture important commonalities. I have further positioned myself with regard to the implicit/explicit vs. intentional/incidental debate. The current state of SLA research does not allow us to establish whether learning takes place with or without awareness in a specific situation. The design of QuickAssist rests on the assumption that learning can be intentional, but also explicit. Exposing learners to specific language elements or language structures will provide opportunities for them to notice, i.e. start to mentally process them, but will not guarantee

that this will actually occur. How processing continues from the point of initial noticing is explained differently in different frameworks. While the Input-Output-Interaction model described in Gass & Selinker (2008) posits a number of distinct stages, such as hypothesis formation, apperception, uptake, etc., the Automaticity/Automatizaton model (DeKeyser, 2001) conceptualises learning as shifting from a stage where learners use implicit rules (declarative knowledge) to produce or analyse a language phenomenon to a stage where these rules are applied automatically(performative knowledge). That learners will have to attend to a new form in some way or another to start the process, however, appears to be undisputed in any cognitive theory of second language acquisition.

There are strands of SLA research that are particularly interested in how language develops in the mind. Based on the assumption that learners possess an innate language faculty (Universal Grammar, mentioned in chapter 2, researchers try to establish the exact nature of this system by studying how second language learning progresses and what similarities (universals) exist among learners of different languages with different first language backgrounds.

In the cognitive tradition, learners are assigned a more active role in the learning process. While still concentrating on what is happening inside the head of learners, traditional views of passive learners that merely act as receptacles have been questioned. Along with the understanding that learners' attitude, motivation, and involvement are vital factors for successful learning, cognitive linguists have established that learner language is a system in its own right and not only an indicator of learners' deficiency. From this perspective, learners are said to use, alter and refine a system commonly referred to as interlanguage (Gass & Selinker, 2008).

From the interlanguage perspective, learner language appears as systematic. While learners process input in the second language, or about the second language, they form hypotheses about its underlying structure. Further exposure to the second language en-

ables learners to verify or falsify their hypotheses. If a hypothesis is falsified, this will potentially lead to changes or refinements of their hypotheses. In this case, the learner's interlanguage is assumed to change. The language he or she produces from now on, will show evidence of this, as the new hypothesis will materialize in the form of new rules that are used in language production.

In essence, as can be seen from this brief overview of the interlanguage hypothesis, the learning process is clearly much more than a sequence of instances of stimulus and response. Learners interact with their environment – they process input, form or refine intuitions about the underlying structures of the language they are learning, take these assumptions as the basis for their own language production and may receive feedback from their interlocutors that helps them to assess whether their utterances have met the standards of native speakers or were at least comprehensible. The learner here has a far more active role in the learning process than in a behaviourist model or in Krashen's 1982 model which, while realizing the importance of other factors such as motivation, conceives of the learner as fairly passive and capable only of moving from one stage in the acquisition process to the next if provided with comprehensible input (i+1). This account takes into consideration that language acquisition progresses in a certain order, a hypothesis that seems to be borne out by a number of empirical studies (cf. Pienemann, 1998). The learner's active involvement in the learning process is ignored or considered only marginally.

The concept of interlanguage is a useful one, also when it comes to the area of vocabulary acquisition and word formation. As will become clear later, vocabulary acquisition, in contemporary accounts, is seen as developing gradually, words are not learnt as individual chunks. Learners acquire the range of meanings, syntactic use, constraints regarding certain registers, etc. over time. The same goes for the development of word formation rules.

It is beyond the scope of this dissertation to give an overview over different strands of SLA research. I have concentrated on introducing some of the concepts usually used by researchers in the cognitive tradition, as they are of importance when learning individual language elements and rules in a setting where learners work on their own with a text, as is the case with QuickAssist. Other important frameworks that are used in SLA are the functional perspective, where learning a language is viewed as being goal driven in the sense that learners are acquiring a language to be able to accomplish communicative objectives. The sociocultural perspective regards language as a tool that is used by members of a society to achieve certain objectives. In order to become part of a society, learners co-construct the tools necessary to operate in it. Language is only one of the tools that they need. The co-construction of language in the learning process that involves the learner and their interlocutors who assist the learner by "scaffolding", helping them to progress new forms. It is based on the American reception of the Soviet child development researchers L. V. Vygotsky and A. N. Leontiev. This theory has been used widely to study classroom based second language learning and has helped to establish that language is a social phenomenon. Like the sociolinguistic perspective that applies sociolinguistic methods to the analysis of second acquisition process, it has served to highlight the importance of social interaction, the role of culture, gender, class, etc. in language learning. As Mitchell & Myles (2004) point out, however, researchers following either of these traditions, have so far not established a theory of second language learning that provides a detailed account of the mental processes involved.

Recently, some SLA researchers have started of looking at the development of a second language as a complex system.

> Language learning can be viewed as a complex and dynamic process in which various components emerge at various levels, to various degrees, and at various times. Individual differences are a natural consequence of learning within

> such a framework because of the dynamic and multi-faceted nature of the emergent system. Slight differences in the relative rate, strength or timing of the component achievements can result in relatively significant differences between individuals in behavioural outcomes.
>
> Marchman & Thal (2005, p. 150)

Using theoretical frameworks such as construction grammar (Goldberg, 2003), researchers study language acquisition using empirical and cognitive approaches. Tomasello (2003) calls this approach of looking at language acquisition cognitive-functional linguistics or usage-based linguistics. The main difference between this perspective and generative approaches is:

> The grammatical dimension of language is a product of a set of historical and ontogenetic processes referred to collectively as grammaticalization. When human beings use symbols to communicate with one another, stringing them together in to sequences, patters of use emerge and become consolidated into grammatical constructions—for example, the English passive construction, noun phrase construction, or -ed past tense construction. As opposed to conceiving linguistic rules as algebraic procedures for combining words and morphemes that do not themselves contribute to meaning, this approach conceives linguistic constructions as themselves meaningful linguistic symbols—since they are nothing other than patterns in which meaningful linguistic symbols are used in communication.
>
> (Tomasello, 2003, p.5)

From this perspective, learning a language means learning of constructions. Tomasello continues:

> If adult linguistic competence is based, to a much larger degree than previously supposed, on concrete pieces of language and straightforward generalizations across them—with many constructions remaining idiosyncratic and item-base into adulthood—then it is possible that childresn's early language is largely item-based a yet can still construct an adult-like set of grammatical constructions originating within these baby constructions (given several years in which they hear several million adult utterances.
>
> (Tomasello, 2003, p. 6)

While Tomasello focuses on first language acquisition, other researchers (e.g., N. C. Ellis & Larsen-Freeman, 2006) have been promoting this approach for SLA.

Ellis, in his meta study on frequency effects, concludes:

> To the extent that language processing is based on frequency and probabilistic knowledge, language learning is implicit learning. This does NOT deny the importance of noticing (Schmidt, 1993) in the initial registration of a pattern-recognition unit. NOR does it deny a role for explicit instruction. Language acquisition can be speeded by explicit instruction. The last 20 years of empirical investigations into the effectiveness of L2 instruction demonstrate that focused L2 instruction results in large target-oriented gains, that explicit types of instruction are more effective than implicit types, and that the effectiveness of L2 instruction is durable.
>
> (N. C. Ellis, 2002)

The specific language areas that are being addressed by QuickAssist are vocabulary and word formation. The following sections will therefore focus on how word formation and vocabulary acquisition is dealt with within SLA. It will become clear that there are

a number of ways in which students can improve in these areas. Extensive reading is especially suited for independent learning and can be implemented using current computer and internet technologies. A section is dedicated to an overview of pertinent SLA studies in this area and to evaluate the results of these studies. The final section of this chapter will deal with motivation, how it influences the outcome of the learning process and ways of fostering it to contribute to a positive learning experience.

2.4.2 Vocabulary acquisition

Köster (2001) claims that "Wortschatzvermittlung" (i.e. the teaching of vocabulary) neither receives proper attention in research, nor in didactics. Further, lexical errors that can be attributed to this neglect hinder successful communication more seriously than do pronunciation and syntax errors, areas that are more prominent in form-focused instruction. Teaching a foreign language, he argues, should be informed by the insights of cognitive science. Here, vocabulary is assumed to be organised in networks. He claims that there are different networks for L1 and L2, but that acquisition of new material always entails a "look-up" of old information in the L1 network. He also suggests that instructors teaching new vocabulary should take the fact into consideration that there are different learner types, should present related vocabulary to increase the learning success and that learning vocabulary ought to be interesting. Semantisation, the acquisition of a form/meaning pair by the learner, is not a unidirectional process. To fully acquire a new vocabulary item the learner needs to negotiate its meaning with an interlocutor, e.g. the instructor, to clarify issues such as semantic scope, degree of formality, contexts of use and collocations.

This issue will be revisited in chapter 3. There, I will argue that the computer is able to assist the learner in the semantisation process, indeed, can do so more effectively than a human instructor in a classroom context. To provide comprehensible input, using different ways of explaining new material and to explain it multiple times is thought to be

beneficial for the acquisition process. It is also important to raise the learner's awareness to differences in meaning, false friends and polysemy. It is important to learn that often meaning and concepts associated with a word or phrase depend on the social and cultural setting.

Köster only considers the situation of vocabulary teaching in DaF, teaching of vocabulary in general has received a significant amount of attention in recent years. On the one hand, there have been a number of publications that are directed mainly at practitioners (for example: Nation, 2001, 2008; Schmitt, 2000, 1998; Meara & Glyn, 1987).

As this dissertation deals with practical issues of vocabulary acquisition to some extent, it is necessary to comment on Nation (2008). It appears that most of the current publications on the practice of teaching and learning vocabulary in a foreign language, or at least in English, use his guidelines as something like a gold standard. Nation is of the opinion that vocabulary has to be divided up into different categories according to:

a their frequency in representative corpora, and

b vocabulary for specific purposes.

Nation (2008) suggests that ESL courses should aim at teaching the 2000 most frequent word families (see discussion below). Courses preparing learners for university should also cover the 570 word families from the academic word list (Coxhead, 2000) The vocabulary levels test he proposes measures learners' knowledge of samples from the most frequent word families divided up in bands of 1000 words each. He also has specific recommendations on how to integrate vocabulary teaching into lessons, when to teach what vocabulary and how much time to spend on it in lessons. An essential part of vocabulary acquisition is reading. While reading is beneficial to the acquisition of new vocabulary, according to Nation, care has to be taken that the amount of unknown vocabulary has to be kept to a minimum. Following his recommendations, texts adequate for

specific students should contain about ninety-six percent of words that are already known to them. The remaining words can either be ignored, as understanding the text does not depend on knowing them, or can be inferred from the context. I will return to Nation, word frequencies and vocabulary coverage.

On the other hand, there is a large amount of literature available that considers vocabulary acquisition from a psycholinguistic or cognitive point of view (for an overview cf. Singleton, 2000, 1999). A perusal of recent publications in this area will show that the field is currently of considerable interest. The following sections are intended as a brief introduction to some important research questions in the field of vocabulary acquisition.

2.4.3 How many words does a particular language have?

This question is extremely hard to answer for a number of reasons. First of all, the term word is ambiguous, and it is necessary to define precisely, what we mean if we speak of a word. Word could potentially relate to what linguists usually refer to as tokens. The number of tokens in a certain text equals the number of space delimited entities (this is only one way of determining the boundaries of words: we will return to this issue shortly). If certain word forms occur more than once, they still count as individual tokens.

(2.199) This sentence has nine tokens, although has occurs twice.

If multiple occurrences of wordforms are not counted individually, we arrive at the number of different wordforms in a text. This set is called types. Taking into consideration that many wordforms belong to the same inflectional paradigm, and counting only the base forms of these paradigms, we refer to the class of lexemes. Thus, bin, bist, ist, war, waren are individual wordforms, different types, but all belong to the same paradigm which is usually referred to by sein (to be), the infinitive form.

Even stating that words in this context ought to refer to the lexemes, does not solve the problem. It is still necessary to decide whether derivations should be counted separately, or not. For example, do the verb 'schnell'(quick) and the adverb 'schnell'(quickly) count as one or two lexemes, do we consider the preposition 'seit'(since) and the conjunction 'seit'(since) as one or two lexemes? Although the two words arguably look the same, they are formed on the basis of regular derivational processes. Many German adjectives and adverbs look alike, so much so, that some German grammars talk about the adverbial use of adjectives rather than adverbs in some contexts:

> Wir schließen uns dieser Position nicht an [i.e. consider adjectives modifying verbs to be adverbs], sondern plädieren für eine Zuweisung zu den Adjektiven und wollen nur noch von den adverbialen Adjektiven sprechen.
>
> (Eisenberg, 1985, p. 220)

Another question is whether compounds should be considered as one word or a combination of several words. It does not really help to use orthography as a criterion here. Bauer (1998) shows that 'girlfriend' can be found in three different varieties, 'girlfriend', 'girl-friend', and 'girl friend', depending on what dictionary one choses to consult. In German, compounds are usually written together, but the 'Rechtschreib-Duden' (*Duden. Die deutsche Rechtschreibung*, 2009) is painfully aware of the fact that no matter how many spelling reforms and rules exist, there is still ample room for confusion and introduces the pertinent section containing nineteen rules with the following comment:

> Die Unterscheidung von getrennt geschriebenen Wortgruppen und zusammengeschriebenen Zusammensetzungen ist nicht immer eindeutig möglich. Wo die nachstehenden Hinweise und das amtliche Regelwerk keine Klarheit schaffen, sollte sowohl Getrenntschreibung als auch Zusammenschreibung toleriert werden.

(*Duden. Die deutsche Rechtschreibung*, 2009, p. 48)

Questions pertaining to this issue that also need mentioning are whether diachronic aspects should be considered, or whether only the 'contemporary' language is relevant (how can it be defined?); whether to include regional varieties or to postulate a 'standard'; and also what to do with specialized vocabulary and loan words (while members of the medical profession might be expected to know what 'rhinotillexomania' is, the average native speaker is probably unaware that it refers to compulsive nose picking).

A pragmatic approach that can be taken is to count the number of entries in a dictionary. Nation (2008) uses the 118.000 words in the New Webster's Unabridged Dictionary because it is not historical. But even using the number of entries as the basis for an estimate, one will end up with wildly different results depending on which works one consults. Moreover, the criteria for the inclusion of a certain word vary from publisher to publisher, as do the sources editors use to assemble their word lists, which nowadays in general consist of large corpora. While these usually attempt to be balanced, most of them, arguably, consist of written texts that are widely and easily available in electronic form, such as newspaper texts.

Statements regarding the size of a vocabulary of a certain language should always be considered with some caution: The size of the vocabulary of a particular language is regarded by some researchers as a factor that determines how hard or easy it is to become reasonably proficient in it. (Nation & Meara, 2002) claim that the problem with learning English is not so much its morphology, but the size of its vocabulary. This is larger than that of many other languages, they argue, because of the Anglo Saxon, Norman and Greek and Latin influence on English. While it might be relatively easy to learn English well enough to "get by," it will take a long time to become proficient to the extent that the vocabulary items appropriate for a certain context or register are used rather than their synonyms which are inappropriate for the context. Claims about the size

of vocabulary should always be taken with a grain of salt as Nation (2001) shows. If we were to follow this line of argumentation, however, the case of German would probably be very similar. Not only did Greek and Latin have a considerable impact on German, at least in formal registers, German also uses a number of French loan words, and continues to borrow freely from English. Polenz (2000) also lists Spanish, Italian and Hebrew as having influenced the German vocabulary.

Rather than considering the entire inventory of words in a language, it is also possible to focus on the individual speaker.

2.4.4 How many words does the average native speaker know?

Nation (2001) points out that studies trying to determine the average vocabulary until very recently were flawed and results of the individual studies varied considerably. Among the reasons for this discrepancy was the concept of word researchers used in their studies which were usually dictionary based.

A study might be carried out along the following lines: Take a representative sample of entries from a dictionary and test how many of these words a participant 'knows', use the result to calculate an estimate of how many words the participant's vocabulary size comprises.

As mentioned above, it is not easy to define precisely what a word is. In addition, dictionaries vary widely in whether they list inflectional forms of a base form all under one lexical entry or to have several, whether to list derivations together, or individually. Thus the choice of dictionary used in a study, will have an important effect on the outcome of the study. More recent studies use the concept of word families. A word family comprises the base form, inflectional forms and a well defined set of derivations. It is reasonable to assume that derivations based on productive prefixes and affixes are not stored separately.

Thus, words like 'uncover', 'redirect', and 'quietly' will be considered members of the word families 'cover', 'direct' and 'quiet', respectively. In addition, more recent research has started to distinguish compounds that are transparent and can be decoded on the basis of knowing the meaning of the individual components of the compound, and those that are lexicalised or idiomatic (e.g. Anglin et al., 1993).

It is also important to define precisely what is to be considered knowledge of a vocabulary item: Does the participant have to be able to produce the vocabulary item (productive knowledge), or does she have to be able to understand it in written/spoken form in a certain context/by itself (receptive knowledge). Studies aiming to find out what the productive vocabulary size of an individual is will aim at evaluating her active use of vocabulary items. While a large enough record of the language produced by the individual over a certain time will probably reveal most of the high frequency items, the number of middle and low frequency items will be hard to estimate. I will return to the concept of vocabulary items.

Nation (2001) reports of studies that claimed that educated native speakers of English have a vocabulary size of about 155,000 basic and derived words. Using more recent methodologies, researchers have come up with far lower numbers. In another study the size is estimated to be less than 20,000 words (Nation, 2001).

Connected to the question of what the average vocabulary size is, is the question of whether there is a difference across social strata. As recently as 2008 Flynn (2008) claims that pre-school children in professional families are exposed to 2150 words, while those in household depending on welfare only to 650 words. This seems all too familiar and reminds one of the discussion on restricted and elaborate code within early sociolinguistics (Bernstein, 1971).

Geoffrey Nunberg, in a radio feature on NPR on September 3, 2002 said the following, commenting on a similar study:

> [The researchers] did find that the average welfare mother used only about 1000 different words in talking to her kids over the several hours of parents' talk that the researchers recorded. But to put that in perspective, the average professional parents only used about 2000 different words in talking to their kids – and that in considerably larger samples of speech. That scarcely means that the professional parents had 2000-word vocabularies, but only that parents of all classes tend to talk to kids in simple language.

While I do not wish to indicate that linguistic studies should not play a part in improving social disparities, a study of the pertinent literature will reveal that many of the projects attempting to establish the vocabulary size of people in general or specific populations are seriously flawed.

It is interesting, however, especially for second language acquisition and teaching research, to find out what the minimum vocabulary size is that enables us to communicate.

2.4.5 How many words are necessary to communicate?

This, again, is a question that is highly debated. Can any conclusions be drawn from the studies reported above? Does the alleged exposure to merely 650 different words during childhood predispose people with a welfare background to 'make do' with a 'restricted code' of maybe a thousand words?

It is important to consider what exactly is meant by communicating and completing communicative tasks. While most tourists will be able to master most of the communicative tasks during their holidays with a very limited vocabulary, or even using non-verbal communication, learners that take language courses during secondary or post secondary education will usually face more complex tasks. We will return to this issue later, but if the motivation for learning a language in such a context is often connected with the desire

to become “native like” then the answer to this question also depends on the answer to the question above. Provided we define the ability to communicate as correlating with the amount of words the average speaker knows, then the targeted vocabulary size should be as close as possible to that of a native speaker.

Pertinent to the area of L2 learning is also the question of how to establish the vocabulary size of a learner. If we were able to answer the above question and had a method to establish a learners' vocabulary size, we would be able to quantify how far the learner has progressed on the way to being able to function adequately in an L2 communication. Early computers were used to analyze the relative vocabulary size. Meara & Glyn (1987) report on a computer program that presented ESL learners with a list of words and asked them to identify the words they already knew and words that they were not familiar with. What the participants did not know was that fifty percent of the words did not exist in English. Based on the correctly identified 'real' words and the incorrectly recognized nonce words the computer then calculated a score that was used to measure the relative vocabulary size. The underlying mathematics were adapted from a military study that tested naval officers on identifying submarines on a sonar. The ESL learners where first exposed to words that tested their knowledge of the band of the 1000 most frequent words of English. Provided their score was high enough, they moved on to the next level until they failed to provide enough correct answers. They were then tested on five percent of the words in the vocabulary band they failed in order to provide an accurate measure of their vocabulary size.

2.4.6 How many words are necessary to comprehend a text?

One specific case of communication is written communication and the part of it that is of particular interest for this dissertation is reading, which is why it will be dealt with in some detail here.

Grabe (2009) examines the question of how many words have to be known in order to adequately understand a written text. His argument involves frequency counts of words and is largely based on Nation (2001) and Schmitt (2000). According to Schmitt and Nation, the most frequent words in English account for the percentages of words in academic tests indicated in figure 2.2. The statistics are based on the vocabulary coverage of a large corpus of English texts.

Table 2.2: Word frequency coverage of academic texts

the	6–7% coverage
top 10 words	22% coverage
top 50 words	37% coverage
top 100 words	44% coverage
top 1000 word families	71% coverage
top 2000 word families	76% coverage
BNC 3000 word families	86% coverage

Note. From Grabe (2009, p. 270)

A conclusion that one might draw from this is that teaching students the most frequent 3000 words of a language will enable them to adequately understand a text of a general nature. Grabe agrees, however, with Nation who argues that knowing eighty-six percent of words in a text does not suffice to achieve an adequate understanding. Eighty-six percent of coverage would mean that at this level learners will still be unable to understand seven out of every fifty words in a text, which is a considerable number. Moreover, the likelihood that the understanding of the text depends largely on the knowledge of these less frequent items is fairly high. These words are often content words that represent key concepts in the text. Nation estimates, that the proficient native speaker usually knows ninety-eight to ninety-nine percent of the words in text, provided that it is a text that does not involve specialized vocabulary that they are not familiar with. This provides them with enough information to understand the text adequately without knowing the remaining words, or enables them to guess from the context what the meaning of unknown words is,

if they are of key importance. In order to be able to understand a text with some guidance of an instructor, learners, ideally, should be familiar already with at least ninety-fife to ninety-six percent of the vocabulary in a new text. This threshold criterion goes back to Laufer (1992).

Even if one was to accept Nation's recommendation that appears commonsensical at least for English, there is still a problem of applying those insights to other languages like German. The most important problem for learners of German, besides mastering inflectional morphology, is the segmenting of complex words. Measuring vocabulary coverage in English is fairly straight-forward. Statistics like the one in table 2.2 are easy to do for English. A word in a given text is considered covered, if its base form, an inflected form, or a form with a common derivational affix is found in the frequency list. The frequency lists used here do not list compound forms as single items, but rather list the constituents individually if they are separated by white space.

If we use the same system to establish vocabulary coverage in German, a far greater list of frequent words would be needed to achieve similar coverage. The following texts only show the words that are covered by the x most frequent words in a German corpus (*German Word Frequency List*, last accessed: 17 September 2010). Compounds remain unanalysed. All words that are not covered are substituted by blanks in figure 2.6. With the 5000 most frequent words, the list used to determine the vocabulary coverage is far larger than 3000 words in table 2.2, but the coverage is lower than eighty-six percent. It could be argued, however, that the text used in the example contained a number of names and that the algorithm counted all numbers as unknown, and that this has an effect on the result. But the fact remains that the concept of word used in the context of vocabulary coverage studies is problematic if we consider a language like German. For German, it seems to make more sense to not consider words, but morphological constituents, such as free and bound roots, and derivational affixes. Establishing the list of most frequent

morphological constituents and study what their overall coverage of a given text are, would probably yield results similar to Grabe's data for English.

The example also serves to illustrate that the ability to segment complex forms like *Nationaldichter, Mahnbriefe, Rundfunkgebühren*, etc. is important. German native speakers are able to analyse novel words into individual constituents, because their mental lexicon contains information in some form about morphological constituents that are available in German and how these can combine.

```
Considering the 100 frequent German words:
_____ _______ ____ ____ _____ ___ Jahre nach seinem
___ sollte der _______________ _________ ________
_________________ ______ Die ___ ___________
__________ an die ___________ _________ ________
___________ und _______ sich _____ um __________
für die _____ _______ Die ______ waren __________
an _____ _________ ________ und ihre ____________ war
________ ________ _____ ____ _____ _______ zu seinem
_________ __ und ___________ ______ _____ es _____
In der _________ ________ ___________ in _________
________ in _______ _____ man dies _________ für
einen __________ ______ ____ einem _______ an die
_________________________ ___ ____ der ____ ___________
_____________ ____ nicht mehr in der ____ sei ein
_____ __________ soll ein ________ _____________ der
_________________ _______ sein Die ___ ___________
am ________ __________ der ________ __________ die
_________ _____ und ________ sich um eine __________
___ _______ mit einer ___________ an _____ _______
______ nicht ______ __________ würden _____ sei
_________ ________ kein so _____________ ____
sagte eine __________ Die _________ von _________
________ _______ die ___ nach _______ _______ von
einem _____________ für die __________ _________
Tokens: 173.0
Unknown: 83.0
Coverage: 0.520231213873
```

Figure 2.3: German word frequency coverage - using the 100 most frequent words

```
Considering the 500 frequent German words:
_____ _______ ____ ____ Über ___ Jahre nach seinem
___ sollte der _______________ _________ ________
_________________ ______ Die ___ ___________
__________ an die ___________ _________ ________
___________ und _______ sich jetzt um __________
für die _____ _______ Die ______ waren __________
an _____ _________ ________ und ihre ____________ war
deutlich ________ _____ doch _____ Angaben zu seinem
_________ __ und ___________ machen hieß es _____
In der _________ ________ ___________ in _________
________ in _______ hielt man dies zunächst für
einen __________ ______ Doch einem _______ an die
_____________________ ___ dass der ____ ___________
_______________ wohl nicht mehr in der ____ sei ein
_____ __________ soll ein weiteres _____________ der
_____________ _______ sein Die ___ ___________
am Dienstag gegenüber der ________ __________ die
_________ _____ und ________ sich um eine __________
Man _______ mit einer ___________ an _____ _______
______ nicht ______ __________ würden _____ sei
_________ ________ kein so _____________ Name
sagte eine __________ Die _________ von _________
________ erhielt die ___ nach eigenen Angaben von
einem _____________ für die __________ _________
Tokens: 173.0
Unknown: 64.0
Coverage: 0.630057803468
```

Figure 2.4: German word frequency coverage - using the 500 most frequent words

```
Considering the 1000 frequent German words:
Schon _______ Herr ____ Über ___ Jahre nach seinem
Tod sollte der _______________ _________ ________
_________________ ______ Die ___ ___________
__________ an die ___________ _________ ________
___________ und _______ sich jetzt um __________
für die _____ _______ Die ______ waren __________
an _____ _________ ________ und ihre ____________ war
deutlich ________ möge doch _____ Angaben zu seinem
_________ __ und ___________ machen hieß es darin
In der _________ ________ ___________ in _________
________ in _______ hielt man dies zunächst für
einen __________ ______ Doch einem _______ an die
_____________________ ___ dass der ____ ___________
_____________ wohl nicht mehr in der Lage sei ein
_____ __________ soll ein weiteres _____________ der
_____________ _______ sein Die ___ ___________
am Dienstag gegenüber der ________ __________ die
_________ _____ und ________ sich um eine __________
Man arbeite mit einer ___________ an Daten _______
______ nicht sofort __________ würden _____ sei
_________ ________ kein so _____________ Name
sagte eine __________ Die _________ von _________
________ erhielt die ___ nach eigenen Angaben von
einem _____________ für die __________ _________
Tokens: 173.0
Unknown: 55.0
Coverage: 0.682080924855
```

Figure 2.5: German word frequency coverage - using the 1000 most frequent words

```
Considering the 5000 frequent German words:
Schon gezahlt Herr ____ Über ___ Jahre nach seinem
Tod sollte der _______________ Friedrich ________
_______________ zahlen Die ___ ___________
__________ an die ___________ Friedrich ________
___________ und bemüht sich jetzt um Erklärung
für die _____ Dresden Die Briefe waren __________
an Herrn Friedrich ________ und ihre ____________
war deutlich ________ möge doch bitte Angaben zu
seinem aktuellen TV und ___________ machen hieß
es darin In der Friedrich ________ ___________
in _________ ________ in Sachsen hielt man dies
zunächst für einen schlechten ______ Doch einem
Hinweis an die ____________________ ___ dass der
____ ___________ _____________ wohl nicht mehr in
der Lage sei ein Radio __________ soll ein weiteres
_____________ der _____________ gefolgt sein Die
___ bestätigte am Dienstag gegenüber der ________
__________ die _________ _____ und bemühte sich um
eine Erklärung Man arbeite mit einer ___________ an
Daten weshalb Fehler nicht sofort __________ würden
Zudem sei Friedrich ________ kein so ungewöhnlicher
Name sagte eine Sprecherin Die _________ von Friedrich
________ erhielt die ___ nach eigenen Angaben von
einem _______________ für die __________ Haushalte
Tokens: 173.0
Unknown: 31.0
Coverage: 0.820809248555
```

Figure 2.6: German word frequency coverage - using the 5000 most frequent words

```
Considering the 10000 frequent German words:
Schon gezahlt Herr ____ Über ___ Jahre nach seinem
Tod sollte der _______________ Friedrich ________
_________________ zahlen Die ___ verschickte
__________ an die sächsische Friedrich ________
Grundschule und bemüht sich jetzt um Erklärung
für die _____ Dresden Die Briefe waren __________
an Herrn Friedrich ________ und ihre Aufforderung war
deutlich ________ möge doch bitte Angaben zu seinem
aktuellen TV und ___________ machen hieß es darin
In der Friedrich ________ Grundschule in _________
________ in Sachsen hielt man dies zunächst für
einen schlechten ______ Doch einem Hinweis an die
_____________________ ___ dass der ____ verstorbene
_____________ wohl nicht mehr in der Lage sei ein
Radio anzumelden soll ein weiteres ____________ der
_____________ gefolgt sein Die ___ bestätigte
am Dienstag gegenüber der Dresdner __________ die
peinliche _____ und bemühte sich um eine Erklärung
Man arbeite mit einer ___________ an Daten weshalb
Fehler nicht sofort __________ würden Zudem sei
Friedrich ________ kein so ungewöhnlicher Name
sagte eine Sprecherin Die _________ von Friedrich
________ erhielt die ___ nach eigenen Angaben von
einem _____________ für die Zielgruppe Haushalte
Tokens: 173.0
Unknown: 22.0
Coverage: 0.872832369942
```

Figure 2.7: German word frequency coverage - using the 10000 most frequent words

Schon gezahlt Herr Poet

Über 200 Jahre nach seinem Tod sollte der Nationaldichter Friedrich Schiller Rundfunkgebühren zahlen Die GEZ verschickte Mahnbriefe an die sächsische Friedrich Schiller Grundschule und bemüht sich jetzt um Erklärung für die Panne

Dresden Die Briefe waren adressiert an Herrn Friedrich Schiller und ihre Aufforderung war deutlich Schiller möge doch bitte Angaben zu seinem aktuellen TV und Radiokonsum machen hieß es darin

In der Friedrich Schiller Grundschule in Weigsdorf Köblitz in Sachsen hielt man dies zunächst für einen schlechten Scherz Doch einem Hinweis an die Gebühreneinzugszentrale GEZ dass der 1805 verstorbene Nationaldichter wohl nicht mehr in der Lage sei ein Radio anzumelden soll ein weiteres Mahnschreiben der Gebührenfahnder gefolgt sein

Die GEZ bestätigte am Dienstag gegenüber der Dresdner Morgenpost die peinliche Panne und bemühte sich um eine Erklärung Man arbeite mit einer Riesenmenge an Daten weshalb Fehler nicht sofort aufgedeckt würden Zudem sei Friedrich Schiller kein so ungewöhnlicher Name sagte eine Sprecherin

Die Anschrift von Friedrich Schiller erhielt die GEZ nach eigenen Angaben von einem Adressenanbieter für die Zielgruppe Haushalte

Figure 2.8: German word frequency coverage - original text

Convincing as Nation's argument might sound, I would still argue that provided with the right tools motivated learners are able to work with a text independently even if they know less than ninety-eight or even less than ninety-fife percent of the vocabulary in a text. Thinking back to my final exam in Latin that required us to translate a section of a Cicero speech after a two semester intensive course, I remember that I had to check at least thirty-three percent of the words in my dictionary before I could even start putting the pieces together. I still managed to get my translation done and answer the exam questions, as did most others in my course. We only had six months to learn the language and prepare for the exam. Concentrating on grammar, we did not have the time to learn much vocabulary. This is not to say that I would encourage instructors to use the grammar translation method in a modern language class room. Making sure that the amount of new vocabulary a learner has to deal with when reading a new text is manageable is certainly a wise idea, both to process the text at an adequate speed and to make sure that learners stay motivated and are not discouraged by the fact that the time they invest into reading the text does not result in them gaining a proper understanding of what the text is about. But, the ninety-six percent should be considered as a guideline and not as a hard and fast threshold that has to be observed under all circumstances. Learners are very different and, in some cases, understanding a small fraction of a certain text might be enough to manage a certain task without any problems.

Another point worth mentioning here is that the vocabulary size necessary to achieve more than eighty-six percent coverage does not grow linear, but rather exponentially. To add another 1000 or even 5000 words does not result in ninety-five percent coverage.

(Grabe, 2009) remarks:

> It seems that a minimum of 10,000 words (not counting inflectional suffixation distinctions) gives an L2 reader a reasonable chance at understanding an academic text, though not reading the text fluently [..] Also, the earlier

> estimate of 40,000 words for L1 students graduating from secondary school [..] matches well with fluent reading requirements even for L2 learners. In the American Heritage word list [..], the compilers note that 43,831 words provide 99 percent word coverage of most texts. Nation (2001) argues that ninety-eight to ninety-nine percent word knowledge of a text is a common expectation for fluent reading. However, it is not reasonable to expect that L2 students read almost any text they encounter in the L2 with fluency, so the real goal is an L2 vocabulary level anywhere above 10,000 words. With more opportunities for fluent reading practice, a greater number of words will become known.

This should provide the reader with some idea of what vocabulary sizes are currently discussed when it comes to reading comprehension. It is of course not possible to draw precise conclusions regarding the influence of the (receptive) vocabulary size on speaking, listening and writing, but is probably safe to assume that increasing the receptive vocabulary size will also be beneficial in these areas.

2.4.7 What does it mean to know a word?

Especially the studies on the effects of incidental vocabulary acquisition and vocabulary acquisition from extensive reading have made researchers aware that to know a word means far more than being able to identify its meaning(s). Pigada & Schmitt (2006), in their overview over pertinent research find that authors have remarked early that this is only one dimension of vocabulary knowledge:

> [R]eading and vocabulary studies have almost exclusively focused on word meaning in determining vocabulary acquisition. However, it has been acknowledged by a large number of lexically-minded researchers that knowing

> a word involves much more than just understanding its meaning. (Aitchinson, 1994; Laufer, 1997; McCarthy, 1990; Nation, 2001, 1990; Richards, 1976; Schmitt, 2000, 1998)

Nation, drawing on Richards (1976) proposed a list of levels of vocabulary knowledge. Its latest incarnation, cited by Grabe (2009), is published in Nation (2001). The following is an example of a German word is based on Nation (2001, pp. 27–28):

> If we say that we know a word, for example the German word *unterentwickelte*, this implies that we know different things about the word. From the point of view of receptive knowledge and use, knowing the word, for example, *unterentwickelte*(underdeveloped) involves:
>
> - being able to recognize the word when it is heard;
> - being familiar with its written form so that it is recognized when it is met in reading;
> - recognising that it is made up of the parts *unter-*, *entwickel-* and *-t-*, *-e*, and being able to relate these parts of the meaning;
> - knowing that *unterentwickelte* signals a particular meaning;
> - knowing what the word means in the particular context in which it has just occurred;
> - knowing the concept behind the word which will allow understanding in a variety of contexts;
> - knowing that there are related words like *überentwickelt* (overdeveloped), *rückständig* (backward) and *verkümmertl* (rudimentary);
> - being able to recognize that *unterentwickelte* has been used correctly in the sentence in which it occurs;

- being able to recognize that words such as *Länder* and *Regionen* are typical collocations;
- knowing that *unterentwickelte* is not an uncommon word.

From the point of view of productive knowledge and use, knowing the word *unterentwickelte* involves:

- being able to say it with correct pronunciation including stress;
- being able to to write it with correct spelling;
- being able to construct it using the right word parts in their appropriate forms;
- being able to produce the word to express the meaning [UNTERENT-WICKELT];
- being able to produce the word in different contexts to express the range of meanings of *unterentwickelte*;
- being able to produce synonyms and opposites for *unterentwicketle*;
- being able to use the word correctly in an original sentence;
- being able to produce words that commonly occur with it;
- being able to decide to use or not use the word to suit the degree of formality of the situation (At present *unterentwickelte* is more acceptable than *rückständige* which carries a negative meaning).

Combining receptive and productive aspects, the following list evolves:

1. Orthography (spelling)
2. Morphology (word family relations)

3. Parts of speech

4. Pronunciation

5. Meanings (referential range, variant meanings, homonyms)

6. Collocations (what words very commonly go with a word)

7. Meaning associations (topical links, synonyms, antonyms, hyponyms)

8. Specific uses (technical, common)

9. Register (power, politeness, disciplinary domain, formality, slang, dialect form

Source: (Grabe, 2009, p. 267)

In the section on extensive reading, I will consider some of the implications of this analysis of vocabulary knowledge on independent reading. It is worth pointing out already, nevertheless, that students are unlikely to gain this gamut of information incidentally in an average language course. Moreover, commonly available vocabulary drill software usually only addresses one or two of these areas, but disregards other areas completely. As will become clear in the chapter on development and when discussing the case study, QuickAssist, while certainly not addressing each of these points extensively, affords learners the opportunity to extend their knowledge on some of the key areas of vocabulary knowledge mentioned above.

But most importantly, for the purpose of this dissertation a fine grained analysis of the concept of vocabulary knowledge leads to the conclusion that the knowledge of the inventory of derivational morphology and its underlying rules are as central to "knowing a word" as are its meaning, uses, constraints imposed by the register, situational contexts, and other factors.

2.4.8 Are all words equally hard or easy to learn?

Laufer (1997) writes:

> Features such as irregularity of plural, gender of inanimate nouns, and noun cases make an item more difficult to learn than an item with no such complexity, since the learning load caused by the multiplicity of forms is greater.

This is certainly true for German.

Concerning the question of word length, Laufer (1997) writes that studies regarding the effect of the length of words on the learnability of a lexical item have been inconclusive so far. On the other hand, the ability of a learner to analyse a complex lexical item into its individual parts is considered important. While derivational morphology in English is equally complex to German, and learners of both languages have to learn to identify derivational prefixes and affixes to be able to analyse complex words, learners of German have to cope with an additional hurdle. Compounds in English, for the most part are separated by spaces. Although some compounds can be shown to be listed in dictionaries without intervening spaces (Bauer, 1998), writers of English in general avoid using long compounds without any spaces.

In German, on the other hand, compounds are generally written as one word, only in some circumstances is a writer required, or free to use hyphens in between individual elements of a compound (see the orthographic rules in: *Duden. Die deutsche Rechtschreibung*, 2009). It follows that the skill to segment words correctly will be required from learners more frequently in German than in English.

The following compounds found in Donalies (2007) are not only challenging to segment, but—I would argue—are also harder to learn than simplex items, if only for the reason that they consist of a number of different simplex items. These individual items, as well as the order in which they appear in the compound, have to be learned.

(2.200) Donaudampfschifffahrtsgesellschaftskapitänskajütentürschlüssel

(2.201) Rhabarberbarbarabarbarbarenbärtebarbier

While these examples are rather extreme, the task of segmenting is non-trivial in many cases, especially for learners of German, and in some cases readers will also have to deal with ambiguities.

(2.202) *Drucker* - *zeugnis*
printer - certificate
print press operator's diploma

(2.203) *Druck* - *erzeugnis*
print - product
print publication or stationery

Although inflectional morphology is not dealt with in this dissertation, mainly because—contrary to derivational morphology—it is usually dealt with adequately in textbooks and courseware, it does seem appropriate to point out that German inflectional morphology holds a number of challenges for learners. German has a very rich morphology compared to English both in inflectional and derivational terms (for details, cf. Eisenberg, 1998; Fleischer & Barz, 2007; Römer, 2006; Simmler, 1998). Besides a complex inventory of inflectional affixes, German also has an elaborate case system and every noun in German belongs to one of three grammatical genders. Forming the plurals of German nouns can result in stem changes (for example,*der Apfel, die Äpfel*(the apple, the apples) similar to some English nouns (for example, *goose, geese*). But, while English has relatively few of those nouns, German has a fair number of them. While most English nouns form their plurals regularly with-*s*, or-*es*, German has a number of different ways to form the plural. In many cases, learners will just have to learn the plural form together with the singular form.

German verbs, similar to English verbs which can be irregular, or regular, are either weak, strong, mixed, or irregular. Strong verbs use different allomorphs for their stems, instead of using a single form. There are different inflectional paradigms (Ablautreihen) for strong German verbs. Learners of German will either need to learn what paradigm a certain verb belongs to, or several forms of every strong verb in order to reproduce its entire paradigm. While the English present perfect is always formed with *have* and the past participle, the German *Perfekt* is formed with a form of *haben* or *sein* and the past participle. Although the use of *haben* or *sein* is to a large extent rule governed, learners still have to learn these rules and their exceptions.

2.4.9 Effective ways to extend the vocabulary range

The question of the most effective way to learn new vocabulary will most likely receive a different answer depending on whom you ask. Not the least important reason for this is an ever growing industry of learning products, and the effectiveness of vocabulary learning is one of the selling points.

Schmitt (1997) in his survey over pertinent literature writes that the effectiveness of individual methods is still under researched. He continues to state that while a variety of methods have been developed and promoted within different language teaching theories, it seems that rote learning and memorization remain the most effective methods, at least when looking at short-term achievements. This view is corroborated by Nation (2001).

Ultimately, though, the answer is likely to be slightly more complex. Schmitt, citing a number of studies, argues that the effectiveness of a vocabulary learning method depends among other things on the degree of proficiency of the learners, their cultural background, their motivation and metacognitive strategies and their belief into the effectiveness of a certain method. While beginning learners seem to generally achieve good results with word lists and rote learning, advanced learners appear to benefit and are more likely to

use imaging and other learning strategies that usually include a deeper analysis of the vocabulary items to be learnt than word lists do.

It has been shown that learners of different cultural backgrounds perform differently when taught and asked to use a new method. Schmitt (1997) reports on an experiment involving Hispanic and Japanese learners of English. While the Hispanic learners perform relatively well using an association technique they are taught, the Japanese learners refused to use the technique and performed less well.

Learning strategies can be categorized using different taxonomies. The most elaborate one was proposed by Oxford (1990). She groups strategies in the following four groups:

- Social: all actions that involve the interaction with others to achieve the goal of learning fall into this category. This can be as simple as asking an instructor for a translation of an unknown word or it can involve group activities like producing diagrams or mind maps.

- Memory: all actions that are aimed at adding new information to memory. This category encompasses rote learning and word lists.

- Cognitive: methods that involve more complex cognitive processes like associations, the use of anchor words, semantic grouping etc. fall in this category.

- Metacognitive: Methods that learners use because they are aware of cognitive processes are categorized as metacognitive. Examples include learning in intervals, grouping words according to certain features, or being selective about what to learn.

Regarding the success of individual learners, research seems to indicate that an important factor is a learner's metacognitive skills. Learners who are aware of what learning methods works best for them, who strategically use them, who are able to self-motivate

etc. seem to outperform those who have a smaller inventory of metacognitive skills. More recent accounts of such studies can be found in White (2008).

A method that according to Schmitt (1997) became popular with the rise of the communicative approach is "guessing from context". This can refer to extra textual information like pictures, gestures, etc., but is usually used to refer to the guessing of unknown vocabulary items in a text by studying the context and trying to infer the meaning from it. The success rate of this approach has been discussed earlier when Nation's threshold level was considered. An area in which learners often have to resort to this method is extensive reading which I turn to in the next section.

2.4.10 Extensive reading

Extensive reading is considered an effective method to extend foreign language learners receptive and productive vocabulary(Nation, 2008, 2001, 1990). Schmitt (2000) writes:

> Reading is an important part of all but the most elementary vocabulary programs. [...] Written discourse [...] tends to use a wide variety of vocabulary, making it a better resource [than spoken discourse] for acquiring a broader range of words. [...] However, most studies show that the vocabulary uptake from reading is really rather small and it is only through numerous repeated exposures from a great deal of reading that any significant number of words are learned [...] What is really needed is extensive reading.

There are a number of factors, however, that have an influence on how successful the learner will be in reading a text in a foreign language. It has been shown that reading proficiency in the L1 has an important impact on the reading proficiency in the L2 (see: Grabe, 2009). Those learners who read a lot in their first language and have a wide range

of reading techniques at their disposal will be able to apply those in an equally effective manner to L2 texts.

It is obvious, of course, that an important factor for the successful understanding of the L2 text is the amount of vocabulary that the learner is unfamiliar, or not sufficiently familiar with. The threshold level of ninety-five to ninety-eight percent with regard to vocabulary items that have to be familiar with has been discussed above. Reading in a second language and the growth of the L2 vocabulary range have been shown to be interrelated. Stanovich (1986, 2000) speaks of a "reciprocal causal relationship between reading and vocabulary". If learners possess an extensive vocabulary range their reading skills are more advanced, and if they read extensively, their vocabulary has been shown to grow. For a more extensive list of research corroborating these results, cf (Grabe, 2009).

Laufer (1997) points out factors that have to be considered with regard to the process-ability of vocabulary items. These include:

- pronounceability
- orthography
- length
- morphology
- synformy(the similarity of lexical forms)
- part of speech
- semantic features
- register
- idiomaticity

Especially length and morphology, I would argue, represent challenging areas for learners of German.

In the following I present arguments for extensive reading as a suitable instrument to maintain and extend the L2 vocabulary, as well as increasing the proficiency with regard to other areas of language.

Grabe (2009) in evaluating Nation (2001) and other research findings presents the quantitative aspect as follows:

> If students read approximately a million words of running text in a year, and if they know ninety-six to ninety-eight percent of the words, they will be exposed to 20,000 to 40,000 new words. (One million words equals ten to twelve shorter novels, twenty-five Newsweek magazines, or sixty-five graded readers.) If a student reads 100 wpm for fourty-five minutes per day, and for 222 days in the year, that student would read just under one million words a year. If students learn one word in ten through context, they will learn somewhere between 2,000 and 4,000 through extensive reading in a year.

These numbers rest on the premise that sufficient exposure to a new vocabulary item in a familiar enough context will increase the likelihood of its acquisition. Gass & Selinker (2008) speak of the saliency of language items. While a certain word might be noticed because of its prominent position, length, or any other feature and be subjected to processing, it might also be the case that only sufficient frequent exposure eventually leads to linguistic processing and to the acquisition. While implicit teaching always attempts to raise the saliency of a language item that is at the centre of a particular unit, it cannot guarantee that learners process it at this point. Moreover, implicit teaching can never hope to cover everything. Nation (2008) argues that foreign language instruction can only aim at teaching learners the most frequent vocabulary terms. Words occurring only infrequently

do not warrant that they be dealt with at length in class. Learners will have to learn them on their own.

Following Nation's (2008) definition of what it entails to know a word, it seems obvious that a complete understanding of a word can only be achieved by being exposed to it in different contexts, with different neighbours, used by different speakers and in a variety of different registers, if applicable. While rote learning vocabulary lists clearly have the advantage that vocabulary items can be memorized fairly efficiently, at least short and mid-term, the range of information that learners memorize together with a certain form, for practical reasons, has to be very limited. Learners do not normally attempt to memorize dictionary definitions which generally aim at being comprehensive. Usually learners memorize a form and one or two translations, they might also memorize some common collocations and details on its inflectional properties if they are marked, but cannot hope to cover everything there is to know about a word. This kind of knowledge can only develop gradually over time. Extensive reading will have the effect that learners keep meeting a word in different context and eventually develop a more complex understanding of a word.

Learners have very different interests and motives for learning a language as we will see in the next section. Since language teachers cannot hope to teach every learner exactly the vocabulary they need for their specific purposes, unless the learning setting is of a one one one nature, learners will have to learn vast portions of the vocabulary they need on their own. To do so, one of the best ways is to read authentic texts from their fields of interest in the target language. Only here will they meet the vocabulary they have to acquire used in a way that is adequate for the particular field.

In my opinion, extensive reading is an important way to extend one's vocabulary. At an advanced language acquisition stage, it might well be the only one. But extensive reading does not only increase the vocabulary. Readers are exposed to a wide variety of

different constructions, idioms and sentence patterns. By the same token that researchers claim learners acquire new vocabulary, they will also acquire knowledge in the area of morphology, syntax, semantics and pragmatics of a language. As Grabe (2009) points out, reading is also a skill that we constantly refine in order to achieve greater speeds, faster or more comprehensive understanding of texts, the ability to read selectively, skim, skip, etc. All these are skills that are important beyond second language acquisition and appear well worth practising. QuickAssist provides a platform for second language learners to practise extensive reading in German and study different aspects of words if they chose to do so.

2.4.11 Motivation

The outcome of learning is usually seen as being determined by the learner's motivation. What motivation actually refers to, however, is often not made explicit. In this section, the concept of motivation is briefly discussed and its importance for language learning in particular will be highlighted.

Motivation in the context of language learning was first researched by a group of Canadian researchers in the early 1970s. Gardner & Lambert (1972) evaluating a study with university students learning French in Ontario claimed that there are two distinct kinds of motivation:

- Integrative motivation according to Gardner is the stronger form of both kinds of motivation and results in better learning results. Integrative Motivation is displayed by those learners who have the long term goal of being integrated into the linguistic group whose language they are studying. Learners who express the wish to be able to fluently communicate with native speakers, wish to live in the other country, become native like themselves, etc. are considered to have a high degree of integrative

motivation.

- Instrumental motivation is defined by Garder as the motivation that learners have that express more practical and short terms goals. Examples of learners that Garder would characterize as learners with an instrumental motivation include those that say that they learn a language to pass an exam, because knowing the language will qualify them for a certain job, because they need a working knowledge of the target language for their travels, etc.

Both Dörnyei & Skehan (2003) and R. Ellis (2008) point out that the terms used by Gardner & Lambert were criticised in literature because of the way they are defined (the definition in which integrativeness depends on integrative motivation and integrative orientation and vize versa, seems tautological to be sure). They also point to a number of articles that question the validity of Gardner's claims with regard to which of the two kinds of motivations will yield a better learning outcome.

More recent approaches to motivation in second language learning have included other factors that contribute to motivation, expanding on the narrow intrinsic factors promoted by Garder. Researchers (cf. Dörnyei & Skehan, 2003) have proposed that models of motivation have to be more comprehensive and include among others the following:

- general motives regarding L2 related values
- learners' self-confidence and self-esteem
- class room environment: personality of teacher, peer groups
- curriculum and teaching material
- distractions

Moreover, while motivation was conceived of as being a stable variable in earlier accounts, more recent literature (Dörnyei & Schmidt, 2001) considers motivation a dynamic system. Learners motivation changes from day to day and is influenced by a variety of different factors including the ones in the above list. R. Ellis (2008) also points out that learners are able to control their motivation. The term self-regulation refers to "the ability to monitor one's learning and make changes to the strategies that one employs"(R. Ellis, 2008, p. 687).

One of the assumptions that underlie the design of QuickAssist is that learners' motivation can be increased and sustained by not only enabling them to process a text in German more rapidly than they might be able to using traditional references, they are also able to decide on their reading material themselves. The ability to look up words in different contexts, find out about their internal structure, etc. is intended to motivate them to explore a range different aspects of certain words and prepare them for independent research while at the same time providing them with quick results that enables them to complete the current task of understanding a particular text.

2.5 Theory and practice

In this section, I want to conclude this chapter with a few remarks regarding the role of vocabulary acquisition and teaching and learning about German word formation within German language didactics and pedagogy. This chapter has shown that an extensive vocabulary is an important aspect of language proficiency. It has been shown that knowing a word is a concept that subsumes a number of different knowledges. Meaning is only one, albeit important, aspect. It is equally important to know about its form, pragmatic function, cultural contexts and other aspects. In this section, we will look at how DaF, instructors, curriculum planners, and textbook authors deal with the area of vocabulary

acquisition. The case of CALL software and German CALL software in particular will be discussed in chapter 3.

2.5.1 Vocabulary acquisition: theory and practice

Zimmermann (1997), in her analysis of the importance of vocabulary instruction from the grammar translation method days to the present, notes that while it is a commonplace that vocabulary is central to mastering a language, no language teaching paradigm so far has given it the attention it deserves. While the grammar translation method put syntactic structures and their thorough analysis in the foreground of instruction and treated only inflectional paradigms and the etymology of words in some detail, later methods placed a large emphasis in correct pronunciation, the learning of phrases and communicative skills. The acquisition of vocabulary "has not been a priority in second language acquisition research and methodology" (p. 17).

It would be natural to expect that German as a foreign language pedagogy should take insights of SLA and vocabulary acquisition research in particular into consideration, that instructors would spend an adequate amount of time and effort on teaching learners German vocabulary and the morphological inventory used in German word formation and its regularities. But there is an academic dispute over whether to include form-focused elements in the curriculum and if so which and how many. This dispute cannot be recapitulated here (for examples of this dispute in the area of German as a foreign language see Rall, 2001).

Research literature suggests that the formal elements that are covered in the language classes are rather restricted. Köster (2001) claims that the teaching of vocabulary does neither receive proper attention in research nor in didactics and that lexical errors that can be attributed to this neglect hinder successful communication more seriously than do pronunciation and syntax errors, areas that are more prominent in form-focused instruction.

He also points out that the notion of a basic core vocabulary in the vein of Ogden's Basic English (Ogden, 1930) is no longer feasible in a modern world which is constantly in flux, and where there is a demand for specialised and individualised instruction. Lessons can only introduce a limited amount of vocabulary. According to his article the average lesson introduces four new vocabulary items. It is also a fact that courses will always be geared to some degree to a general audience, thus ignoring the needs of the individual student.

This forces the learner to become more reliant on self-study and acquire large portions of vocabulary independently. Classroom instruction should concentrate on providing learners with the communicative practise they cannot get on their own and on teaching them the skills they need in order to successfully assume responsibility for parts of their learning process.

The situation for the teaching of word formation is even more unsatisfactory. Nation (2001) reports on studies that looked at the development of native speakers. While vocabulary growth is persistent and high over the first few years of first language acquisition, there is a point at which the vocabulary growth virtually stops. At this point in the acquisition process, learners develop analytical faculties that enable them to process morphological features of multimorphemic words. After this stage of the acquisition process is completed, vocabulary growth sets in again, albeit at, apparently, a reduced rate. If second language acquisition is in some way related to first language acquisition, and the jury is still out to decide this, then it might prove helpful for learners to provide them with tools that can give them an insight into the morphological make-up of words.

The section on German word formation illustrates that this is a complex, but nevertheless rule governed area of language that learners have to be aware of. Singleton (1999) argues that learners confronted with the problem of communicating something in a foreign language for which they lack vital vocabulary will resort to the strategy of lexical

innovation. To teach learners the laws of German word formation and to provide them with an opportunity to test their hypotheses about how unknown vocabulary items might be derived from known elements will provide them with the necessary strategies to master such a problem. Rings (2001), however, finds no German textbooks or business German textbooks that deal with the subject in more than a cursory fashion. Max Möller (personal communication, 6 April 2008) reports that there is hardly any literature on German word formation that is suitable for German instructors and that German textbooks he and his students analysed in a seminar at the University of Berlin had little material on the subject and many of the pertinent exercises were found to be of inferior quality.

While Olejarka (2008) discusses some new materials available for DaF, these only cover the formation of verbs and are largely concerned with inflectional paradigms. Derivation and compounding are not dealt with in any detail. Thus there seems to be definitely a need for teaching materials that cover German word formation processes in sufficient detail. This need, to the best of my knowledge, has not been addressed in recent years.

On the other hand, recent publications such as Römer (2006), Donalies (2007), and just recently Römer & Matzke (2010) which present German morphology and word formation in a concise and accessible way and address undergraduate linguistics students as well as language teachers are a promising sign. A generation of teachers who have been made aware of the importance of word formation and are equipped with the background knowledge, hopefully, will eventually be equipped with textbooks that cover word formation and other important areas of the German language in sufficient detail and more adequately than current materials.

It has been argued earlier that one way of increasing learners' vocabulary is extensive reading. The next section will look at what materials textbooks used in DaF courses offer learners to do extensive reading.

2.5.2 Extensive Reading

If one considers German text books that are currently used in Canada, it is clear that the reading materials most books offer are not providing enough opportunities for learners to do extensive reading. *Passwort Deutsch* (Albrecht et al., 2008), a book usually used in German as a second language courses in German speaking countries offers reading texts that – even at the most advanced level – hardly exceed 200 words. Books geared more specifically toward the needs of North American students do not fair much better. Texts here usually do not cover more than one text page (for examples, see Moeller et al., 2005; Lovik et al., 2007).

Graded readers are certainly an option for language instructors who want to offer their students more opportunities for extensive reading. From a pragmatic perspective, however, using them means additional costs for course materials. If instructors were to follow Nation and Wang's 1999 advice to use at least one graded reader every two weeks, costs would soon be getting prohibitively high. Moreover, the number of graded readers available for German as a foreign language is far smaller than the ones available for English or French.

2.5.3 Conclusion

In general, then, we can conclude that the acquisition of new vocabulary, including word formation rules does not seem to be an area that is adequately addressed in textbooks and in teacher education. We are probably also able to conclude that this has an effect on the average German course and that many DaF students do not receive adequate instruction in this area or learn enough vocabulary. The same can be said about extensive reading. Textbooks, in general, do not provide enough reading opportunities, graded readers are not readily available or not used extensively enough. Unless the majority of German

instructors prepares adequate reading material for their students themselves, there seems to be room for considerable improvements.

The discussion of SLA in general and vocabulary acquisition in particular showed that competing theories exist within SLA. It is beyond the scope of this dissertation to adequately discuss different language teaching paradigms, e.g., the grammar translation method, the audio lingual method, the communicative approach or the post communicative approach. It is important to understand that each of the underlying theories, e.g., behaviorism, constructivism or socio-cultural theory, makes different assumptions on the learner's role in the learning process, the function of the instructor, the setting in which learning takes place, and the importance of motivation and other factors for learning outcome.

To position myself within all of these different discourses, and without subscribing to any theory or paradigm in particular, I want to list a number of key points that are part of my beliefs about learning and teaching a language, that determine my teaching philosophy and that have influenced not only the design of QuickAssist, but also the structure and content of this dissertation.

- Learning a foreign or second language is a process that involves the active participation of the learner.
- External factors such as motivation, learning strategies and opportunities for hypothesis testing play a vital role in successful language learning.
- If learning takes place in an institutional setting, successful learning depends largely on students' motivation, their ability to use learning strategies and eagerness to actively test their hypotheses. How these qualities are fostered by the learning environment will have a significant influence on the learning outcome.

- One of the key areas that language instruction should encompass is teaching learners to gradually assume increasing degrees of responsibility for their learning progress.

- This includes that learners learn to use tools such as a dictionary, a grammar, CALL applications (and QuickAssist for that matter) in an efficient way and adapt them to their learner type, their learning style and the task at hand.

The design of QuickAssist rests on the following assumptions regarding vocabulary acquisition:

- Learning vocabulary is an incremental process. Rather than claiming that it is merely a matter of exposing learners to vocabulary items a certain number of times in order to guarantee learning, I believe that learning occurs in stages at which the different aspects of knowledge about a vocabulary item is gradually acquired.

- The theoretical framework that accounts for such gradual acquisition over time and that seems to be most closely related to my view about vocabulary acquisition is the emergent field of construction grammar and especially the works by researchers such as Tomasello (2003) grounded in empirical cognitive science.

- As the knowledge about a vocabulary item comprises many levels, a variety of tools providing learners with different kinds of information on the word (e.g., context, morphological structure, synonyms, frequency) is necessary.

- There are different learner types which will benefit to varying degrees from different tools. Rather than making a decision for them and offer them a specific set of tools deemed most suitable for all learners, QuickAssist offers them a range of tools. This enables learners to try out and find the tools which are most suitable for them at a

certain stage and in a given situation. As will become clear later on, of course, also technical issues have had an influence on what tools are provided by QuickAssist.

In chapter 3 QuickAssist is introduced, a CALL application that enables learners of German to work with authentic German texts of their own choice. The application offers a range of functions to users that are intended to assist them with their reading. Before the application is discussed in detail, the chapter will discuss some aspects of CALL in general.

Chapter 3

Computer Assisted Language Learning

The computer has been used in foreign language learning ever since it became widely accessible. For an overview of the history of CALL cf. Heift & Schulze (2007); Levy (1997); Nerbonne (2003). One important point is that the computer can function as a tool that can be used like any other media in the learning process. It does have several advantages over other media that have a longer tradition in the foreign language class room, though. Not only can it substitute all other media: with the proper software applications, it provides the student with a potentially unlimited supply of exercises, with tools for self-evaluation and suggestions for improvement, its patience is unlimited and it will dutifully correct the same mistakes.

In section 3.1, I will revisit the topic of research paradigms in CALL that are already discussed to some extent in the introduction. I will argue that the development of CALL software can be considered as a strand of CALL research if it serves to test hypotheses on certain aspects of CALL.

The following section is concerned with the role of the computer and the learner in CALL. It briefly deals with learner independence in CALL, a topic that has received considerable interest over the last few years.

In section 3.3 the application of natural language processing technologies will be discussed in some detail. Since some of them are used in QuickAssist, it seems appropriate to introduce these technologies, explain how they work and what their applications in the area of CALL can be. I will cover tokenizers, lemmatizers, morphological analysers, part of speech taggers, parsers and corpora. Other NLP technologies, for example the ones that are concerned with the processing of oral language will be ignored since they were not used for the research project.

As mentioned, what follows is a discussion of research paradigms that exist within CALL in some detail.

3.1 Theory and practice in CALL

In this section I would like to provide a short overview of research paradigms currently used within the CALL community. Significantly, CALL is a very diverse field, and positioning myself clearly right from the start is done to make transparent why certain theoretical aspects seem to take centre stage in this text, while others are slightly mentioned or even ignored. As Hubbard (2009) points out in his preface to the latest and most comprehensive overview of the field of CALL, the discipline comprises research, practise and development. He also says that CALL is both exciting and frustrating, naming the diversity of the field as one of the examples for both his excitement and his frustration. It is impossible to provide a comprehensive list of theoretical frameworks and research methodologies that have been historically used in CALL. In short, CALL is an area of interdisciplinary study with contributions made by researchers from such diverse disciplines as education, linguistics, second language acquisition studies, computer science, computational linguistics, sociology, psychology, and philosophy. This list is not intended to be exhaustive (see Colpaert (2004) for a detailed discussion). Not only do all of these disci-

plines have very different approaches to research; research paradigms, schools of thought and theoretical assumptions can vary greatly even within the individual disciplines. Attempting to do justice to this diversity and cover them all is not possible here. I would like to make clear my position within the theoretical discussion.

First, let us return to Hubbard's remark on what the different aspects of CALL are. CALL practice used to refer to the use of computers in second language teaching. Currently, things aren't quite as straight forward any more, since computers have become part of almost everyone's life, not only in the form of the object taking up a sizeable amount of space on our desks. Our TVs, DVD players, our music players, cellphones, etc. represent computer systems or are likely to include them in one form or another. In turn, people have started to use some of these devices in language teaching, especially the use of Apple's iPhone™ and other mobile devices which have attracted the attention of many CALL researchers and practitioners. Chinnery (2006) uses the term MALL (Mobile Assisted Language Learning) to refer to this phenomenon. Nevertheless, we will probably want to distinguish the use of a PC, or a mobile device in language teaching from the use of a DVD player. CALL is not considered CALL because we are using a device that operates with the help of a microprocessor. An important aspect of CALL is that the user is interacting with the machine in some meaningful way. In addition to this, users are not only using a single device anymore. The internet has been gaining more and more importance in CALL over the last ten years or so. This has in effect led to the point that learners do not solely interact with a computer anymore. In many cases they are interacting with other humans using a computer as a means to communicate synchronously (e.g.: in a chat room or "skyping") or asynchronously (e.g.: via e-mail, or posts in forums). The body of research done on computer mediated communication (CMC) is rapidly growing. There does not seem to be a foreseeable point in the future in which the term CALL is as redundant as BALL – book assisted language learning – (Bax, 2003). While certain

technologies will without doubt at some stage be considered as normal in language teaching/learning as is a book and a pen, the fact that computer technologies continue to evolve provides a steady influx of new technologies, new media, and new possibilities for language teaching, and there will always be people interested in finding applications for new media.

Just to illustrate this point: We have seen a fair bit of web-based communication and collaboration tools evolve over the last ten years and all of them, I would argue, have found applications in language teaching somewhere. Before most of us (at the time of writing) have even developed some understanding of what Google's next big project, Google Wave, is about, there is already an animated discussion on some of the CALL mailing lists about how this technology could potentially be exploited for language teaching purposes. Again, I would question whether it is a real novelty at all.

Not only has there been a change of the media and technology that has been studied within CALL, the research questions asked in CALL studies have changed as well, over time. While there are still some studies that compare CALL based language learning to traditional learning, these kinds of studies were certainly more frequent in the past and have been fairly inconclusive so far (for a detailed meta study see Felix, 2005). These so called efficacy studies have been criticised for being reductionist by selecting just one variable, the technology used or not used, and by ignoring all other variables in their comparison. Since learning and teaching is a multivariate and dynamic system, this reductionist approach often fails. Researchers have come to understand that comparing the learning outcome of CALL to traditional learning in quantitative terms is only one of many questions that CALL research can study. There has been an increased interest in questions that are concerned with the uniqueness of the interaction between humans and the computer, or between humans mediated by a computer in a language learning context. As pointed out, it is important what the question is and who asks the question. This will

ultimately determine the research paradigm and the theoretical assumptions made by the researcher. While there have been attempts to streamline research paradigms and offer a common approach to CALL research (for example, Chapelle, 1997; Levy, 2001) they have not had the effect of impinging on the diversity of the field.

The third aspect Hubbard points out is the development. I will argue in chapter 4 in detail for the importance of researchers who are involved in or assume control over the development of a CALL application. First, I would like to give a brief summary of my points here. Currently, the majority of dedicated CALL programs are developed by commercial software publishers who are

- employing professional programmers;
- interested in keeping development costs down;
- interested in a high profit margins;
- interested in keeping their trade secrets.

This has some very important consequences for the quality of the software and for the field of CALL altogether:

- Although language specialists and teachers may be consulted during the development, the development is conducted by programmers.
- The publisher will usually fall back on technologies that are considered tried and true, rather than invest into novel research or adapting "bleeding edge" technology.
- If novel language technology happens to be developed for a new program it will not be available to be reused and improved by others.

In order to develop novel language learning technology, it is important to have access to the insights not only of computer science, but also to those of second language acquisition studies and language teaching practice. In fact, most of the early software that was developed for the first personal computers in the 1980s was developed by language teachers turned hobby programmers (Heift & Schulze, 2007). A similar stance is also taken in (Levy, 1999) who distinguishes CALL design intended to be used in the real language class room and CALL design in which

> the CALL program functions as a testbed for research and is aimed at substantiating a theory, usually a theory relating to an aspect of second language acquisition.
>
> (Levy, 1999, p. 90)

Given this context, I hope to find out whether, by using state of the art natural language processing (NLP) software, it is possible to develop a tool for intermediate and advanced learners of German to help them with vocabulary acquisition and to get a deeper understanding of the rules governing German word formation. Given my background, both as a linguist and a language teacher, the development of the software was to be guided by the insights of SLA and by plausible models of German word formation that are taken from theoretical linguistics. The latter field is not only relevant to this text because theoretical linguistics has always had a historical influence on SLA and language teaching, historically, but more importantly because these models have been used within computational linguistics to develop NLP software. I will return to this subject in section 3.3.

3.2 The role of computers in CALL

The role of the computer in CALL has traditionally been described with the help of dichotomies. In terms of mediation, it can either act as a tool or tutor (using Levi's 1997 terminology). The magister/pedagogue distinction, used by Higgins (1988) is used to assess the degree of control assigned to the computer and the degree of freedom the user is given to make important decisions about the learning process. In a fine grained analysis, these dichotomies can be viewed as continua. Commercially available dedicated CALL software by and large can be located at the tutor/magister end of this continuum. CALL, on this side, is machine centred.

The reason for failing to make the shift to student centred CALL cannot be accounted for by technical limitations. On the contrary, if the power to make decisions about the learning process is shifted from the computer to the learner, the necessity to have it act omniscient (all knowing) and human-like disappears. I would venture that some of the reasons for this are:

- As mentioned, software developers, by and large have no language teaching background. It is reasonable to assume that many of them will have antiquated ideas about how language acquisition should work.
- As long as the software based on those ideas finds a market, there is little economic reason to invest in radically different software.

While instructors and students might have had inflated expectations with regard to the capabilities of tutorial CALL in the past, (see Nerbonne, 2003), many teachers have become aware of the limitations of tutorial CALL. Many of them seem to be reluctant to use CALL in their classes, presumably because they feel that these limitations are too severe to warrant the investment of financial and human resources necessary to administer

them. In addition, as pointed out above, the teaching theory underlying dedicated CALL applications seems antiquated. In the following section, I am discussing so-called learner independence in the context of CALL.

3.2.1 Learner Independence

Learner independence, also referred to as learner autonomy (Oxford, 2008), is a widely used buzzword. This has been criticized for a number of reasons. Pennycook (1997) remarks that the notion of independence or autonomy is a western concept. While independence or autonomy have a positive connotation in western society, it might evoke radically different associations in other cultural contexts that place the society above the individual, in which society acts as a sanctuary for the individual. A state of autonomy, of being outside society in such a context may well be considered undesirable. Schmenk (2006) traces the term autonomy back to ancient Greek philosophy were it was used to describe a political state. This meaning was later adapted by Enlightenment philosophers like Immanuel Kant. As language learners because of the role they assume in the context of language instruction are considered to be in need of guidance and support, it would be wrong to call them autonomous, from this perspective. True autonomy in the learning process is not possible. Even if learners are able to decide on what, where, and when to learn they will still need to use resources for their learning process that are provided by others and cannot be determined by learners. Therefore, the terms autonomy and independence will not be used interchangeably here.

Many researchers are aware of the inherent problems of the term independence and try to use it cautiously. Nevertheless, it is still necessary to define clearly what independence means in a certain context. With respect to the domain of learning, White (2008) locates independence on three different levels:

- Context/Setting: Independence can simply mean that learning takes place without a human teacher, but it can also mean that learners have the freedom to make choices, the freedom to select learning opportunities and the freedom to use resources according to their needs.

- Philosophy/Approach: On this level, independence refers to the roles and responsibilities of teachers and learners in the independent learning context. The teacher's role here is to prepare learners to think about their needs. Learners have to develop the ability to look after their own needs.

- Learner Attributes: Learners have to develop the attitudes, beliefs, the knowledge and the strategies to take actions that support their learning process.

Taking a look at tutorial CALL applications that are available today, it is clear that they are all able to adequately function as the e-learning system on the mentor end of the scale described above. As long as they have full control over the learning material and learner input is constrained, as for example in true/false or multiple choice type questions, they will perform well. It is important to note that only some of them will go through the trouble to give learners a comprehensive explanation for the selection of tasks and the order in which they are administered. Ideally, learners should be able to expect more than a summary of the grammar points, covered, topics of the lesson, etc.

Most modern systems will also provide learners with a range of choices and the option to create individual learning programs by enabling them to select learning materials and to decide in which order they are presented.

It could be argued that most systems enable students to create individual vocabulary databases, that complex systems place a large amount of learning elements at learners' disposal, but it is clear that individual needs are hard to predict and that even sophisticated systems would be unable to offer everything. An architect learning Italian in order to read

Italian publications on architecture, for example, might be able to find a lesson on art history or even texts on some architectural monuments, but nothing that would introduce her to the specific terminology used in her field of interest.

There is little that currently available software does in terms of preparing learners to progress to further stages of independence. In order to be able to continue to operate with the limited set of functionality that was discussed above, the system is forced from a technical point of view to keep the learner in a state of dependence. Once learners develop a certain degree of independence, the learning system will either have fulfilled its purpose or become part of a larger pool of resources learners use to proceed in their language acquisition. I would argue that claims made by the producer of tutorial CALL software to provide a comprehensive program that lead the learner to a stage of proficiency past an advanced beginner level are misleading or overstated.

Returning to White's classification of independence, there is another area that developers of tutorial CALL software could take into consideration. While it is clear that CALL applications can not —and will not for some time— replace a human instructor, they can help to develop the kind of skills that White is pointing to. Including exercises that help develop critical thinking, research tasks, lessons on different learning strategies is certainly possible with currently available technologies. Gradually helping beginning learners to become independent from the system while making them aware of its advanced features (such as dictionaries, grammar references, etc.) would not only help learners, but would also ensure that they would continue to use the application, at least some of its components, as a linguistic tool box.

I do not wish to debate that many tutorial CALL applications have proven very useful and are considered a good addition to traditional language learning, but given the fact that both instructors and learners today are able to use a variety of computer applications, know about the capabilities of these applications and about their limitations, users could

use language related applications as tools just like any other tool computers provide. It is illusionary to depict the computer as more than just a useful machine.

Moving to learner centred CALL would mean to provide learners with a set of tools over which they assume control, learn to use them where appropriate and to assess the level of their reliability in certain contexts. Just as most of us have learned to use a search engine, or to make judicious use of the advice our spell checker and grammar checker give us, we can learn to use other computer tools for our language learning process.

QuickAssist is, in effect, a set of NLP tools, and the user study at the end of this dissertation shows that it is possible for users to learn how to to use them for their learning process.

3.3 ICALL

From Heift and Schulze's 2007 overview it becomes clear that NLP can be defined narrowly as Natural Language Understanding and Natural Language Generation. As such, NLP would be considered a branch of artificial intelligence (AI) research. Others would like to define the area more broadly, (Jurafsky & Martin, 2008) simply write:

> The goal of this new field [NLP] is to get computers to perform useful tasks involving human language, tasks like enabling human-machine communication, improving human-human communication, or simply doing useful processing of text or speech.

Following this definition, NLP comprises more than classical AI applications, like parsers and speech synthesis. Lemmatisers, electronic dictionaries, text corpora and other applications – by virtue of processing natural language in some way – are considered NLP. A look at the pertinent literature (see, for example: Manning & Schütze, 1999; Jurafsky

& Martin, 2008) shows that there are numerous applications available in NLP that are working robustly and are used by the research community, as well as in industry.

Most CALL applications, surprisingly, especially the ones that are professionally developed and targeted at the end user do not actually include any components that are developed by NLP researchers (Jager et al., 1998). Parsers and taggers, corpora and other tools, although having been available for some time and having advanced well past the early experimental stages, are not usually integrated into modern language learning software. A notable exception are pronunciation exercises based on the analysis on waveforms which are becoming more popular in modern CALL applications.

(Nerbonne, 2003) comes to the conclusion that CALL, for the most part, uses the following technologies:

- (simple) database technology;
- digital audio and video;
- hypertext;
- network communication.

He also argues that expectations as to what CALL and NLP based CALL can achieve are often inflated. While there are some parser-based CALL applications available that are able to work with learner language to some extent and to provide corrective feedback (Amaral, 2007; Heift, 1998; Nagata, 1992; Schulze, 2001), these are only able to work under fairly restrictive conditions. Parsing free input and providing adequate feedback is not possible.

In the following sections, I will describe some of the NLP technologies that work fairly robustly, that have found a variety of applications in research and in the industry and that can also be used in CALL applications, as the QuickAssist user study will show. The

actual implementation of these modules in QuickAssist will be discussed in the following chapter.

3.3.1 Tokenizers

Tokenizers, in many NLP applications take on the first stage of processing an electronic text. They are used to analyse the text in distinct elements, or tokens. As the processing is concerned with natural language, ideally the resulting tokens should represent what we customary refer to as words. The problematic concept of word was discussed in chapter 2.4. For most western languages, tokenizers use white space and punctuation marks to establish the boundaries of tokens.

Some of the more common problems of this approach, as for example the problem of abbreviations like: e.g., or i.e., or the way numbers are usually written, e.g. 1,000.00, can be dealt with by heuristics using lists of common abbreviations and pattern matching algorithms to establish components of numbers. Compounds in English that are not usually written as one word, on the other hand, represent more complex problems.

In German, it is especially the synthetic tenses and separable verbs that pose a problem for tokenizing.

(3.1) Er *hat* gestern ein neues Auto *gekauft*.

(3.2) Sabine *hatte* eine Menge *vor*.

While in 3.1, *hat gekauft* is a perfective form of *kaufen*, most tokenizers will analyse the two elements as individual tokens. This is due to the relatively free word order we have in German and due to the fact that the synthetic forms use several different auxiliary verbs and the participle forms are not always predictable.

In example 3.2, the verb is a form of *vorhaben*, however, it is split up and the individual elements occur in different positions in the sentence. Using a part of speech tagger (see below), these forms can usually be identified and treated as one token.

A part of speech tagger was used in QuickAssist with some success during the development. Because of memory issues, however, the version used in the study did not use it, in order to minimize possible complications that users could face when installing and using the software. This will be explained in more detail in the next chapter.

Nevertheless, tokenizing western texts is considerably easier than many Asian texts, as the symbols used in Asian writing systems are often not delimited by any whitespace. This necessitates a far more complex analysis of the text to arrive at a useful way of tokenizing it. For details, cf., for example, Jurafsky & Martin (2008).

3.3.2 Lemmatizers

As pointed out in chapter 2.4, it is often necessary to establish whether a word in a text is a member of a certain paradigm. A practical example is a dictionary look-up. As dictionaries usually list the base form of a word together with its definition, or translation, an NLP application has to be able to establish the base form, or lemma of any token that is to construct a query to a dictionary database. In languages with a fairly simple morphology, stemming algorithms are sometimes used to find lexical base forms. In English, for example, plurals and the third person singular are usually formed by adding -(e)s to the base. By providing extra rules that cover exceptional cases such as *ox-oxen* these algorithms perform fairly reliable. Although stochastic models exist for languages with a more complex morphology (cf., for example, Creutz & Lagus, 2002), it is often more reliable to create lists of word form / base form pairs and use these to reliably establish the lemma for any given base form. This method was used in the design of QuickAssist.

The program used to create the word form - base form list that is used in QuickAssist is Morphy. It deals fairly reliably with most common nouns and verbs. Thus, the form *ist* will be identified as the third person singular present tense form of *sein*. This information can then be used, for example to look up the base form in a bilingual dictionary which in return will provide the information that *sein* translates into to *to be*. Articles and demonstrative pronouns, on the other hand, are not converted into base forms, but rather into stems. In these cases, The base forms were added manually to the list. Details are discussed in the following chapter.

3.3.3 Morphological analysers

A morphological analyser is used to establish inflectional and derivational characteristics of a word form. It analyses a form into its individual morphemes. These in return provide the information necessary to determine the part of speech of the form, its position in an inflectional paradigm were appropriate and the individual words or affixes that it consists of. All this information can be used to:

- look up the base form of a word (the lemmatizer can thus be considered a specialized morphological analyser);
- display individual elements of a compound word;
- generate the complete inflectional paradigm a form belongs to;
- identify other elements, such as prefixes and affixes.

A morphological analyser is usually based on finite-state automata as they were developed by Koskenniemi (1983) and are still used extensively in a Finite-State Morphology setting (Karttunen & Beesley, 2003). It attempts to analyse words into morphemes. A finite state automaton analyses a string element by element. If the element in position one matches an

expected result, a condition has been met and the automaton processes the next element. If an element does not match an expected result, the automaton fails. If all elements in a string are matched, it succeeds. Automata can contain loops and can also contain branches where the processing of an element depends on an element that was processed earlier in a string. By combining finite state automata, complex systems can be built that are able to process natural language. Many programming languages use so called regular expressions that are used in pattern matching operation. These regular expressions are usually implemented as finite state automata. For a recent and detailed discussion of finite state automata and morphological analysis using computers, cf.: Roark & Sproat (2007).

As the nature of the morpheme cannot be established only on the basis of a word alone (e. g.: in order to decide whether /-en/ in /formen/ is a plural morpheme attached to a noun base or an affix to a verb base, forming the first, or third person plural present tense form of the verb "formen" or its infinitive), additional information is needed by the analyser. It is usually a Hidden Markov Model (Jurafsky & Martin, 2008) that is used to try to infer the part of speech of the analysed word. I will return to this issue in the next section. Once a word has been analysed it is possible to generate its inflectional paradigm, generate the base form (infinitive, nominative singular, or the positive without inflectional affixes) for a dictionary look-up, to find examples of words with the same or similar characteristics in order to use them in exercises, and to provide information on idiosyncrasies related to the word, one of its constituents or the rules governing word formation with this word or its parts.

There are different approaches to morphological analysis. One important factor is whether the analysis of new forms is done automatically or unsupervised (see for example: Creutz & Lagus (2002)), or supervised.

The number of morphological analysers for German is relatively small. *Word Manager* (Hacken, 2003; Hacken et al., 2006; Hacken & Domenig, 1996; Hacken & Tschichold, 2001) is a fairly sophisticated tool, able to deal with derivation and compounding. Word Manager was created on the basis of complex morphological databases manually created by a team of linguists. It has not been integrated into a CALL application. The developers decided to market the software closed source and sell it commercially. It is possible to query the Word Manager Database manually via the website http://www.canoo.net. While this allows users to access all the information contained in the database, it is not possible to access it via a web service or another interface, so that a software application would be able to directly interact with the database.

The number of ICALL applications that do use a morphological analyser is equally small. Examples include *IDAZKIDE* (described in Ilarraza et al., 1999)) and *Glosser* (described in Dokter et al. (1998)). Neither of these applications was designed for German. *Glosser* uses another NLP technique: corpus alignment. With the help of an algorithm bilingual corpora are automatically aligned using paragraphs and sentence boundaries. Instead of, or in addition to, looking up an unknown word or phrase in a traditional dictionary, an aligned corpus can be searched for occurrences of this item and their translations in the parallel corpus can be inspected to assist in finding the most adequate translation in a given context.

3.3.4 Part of speech (POS) taggers

POS taggers use an algorithm calculating the probabilities of a token being a member of a certain POS, by taking into consideration the tokens preceding and following the token in question. The most probable POS-tag will then be assigned or in case probabilities for different POSs are high enough not to be considered dismissible, the token will be marked as ambiguous. The algorithm underlying this procedure is based on the Marakov

rules and looks at four or more neighbouring words to establish these probabilities.

The probability of a preposition following another preposition is considerably lower then that of an article following a preposition. Apart from the possibility that a token is correctly identified there are three other cases to consider. The tagger is unable to assign a tag at all, a token is ambiguous or it will make a wrong assignment. Even though state-of-the-art taggers tend to be so called hybrid taggers, working both rule and probability based they are still unable to achieve 100% accuracy.

The part of speech of tokens is especially important to establish the syntactic structure of sentences. There are a number of POS taggers available, the ones that are freely available and work with German include the so-called Stanford tagger (*Stanford Tagger*, last accessed: 13 September 2010), developed by Christopher Manning and his collaborators, and TreeTagger (*Stuttgart Tree Tagger*, last accessed: 13 September 2010), developed at the University of Stuttgart.

3.3.5 Parsers

Parsers are used in NLP applications to establish the syntactic structure of the sentences comprising the text that is to be processed. In order to establish this structure automatically, in principle, there are two routes that can be taken. Statistical language processing (cf. Manning & Schütze, 1999) aims at developing mathematical methods to establish, among other things, the structure of sentences. In recent years the reliability of statistical parsers has increased dramatically. Literature also refers to statistical parsing as shallow parsing.

The alternative route an NLP application can take is deep processing. Formal Grammars such as the ones outlined in chapter 2 are used to establish possible syntactic analyses of a sentence. The grammatical framework work most commonly used for syntactic

analysis is HPSG (Pollard & Sag, 1994, 1987; Sag et al., 2003). This theory is based on the assumption that all syntactic units (phrases and lexical items) are headed. The head of a unit determines syntactic, semantic and morphological characteristics of the unit as a whole. All elements can contribute to the characteristics of the unit by propagating certain features to the head. The resulting framework consists of a very small set of rules and a rich lexicon in which all morphological, syntactic and semantic information is stored. This makes it relatively easy to implement it in a computational formalism compared to other frameworks.

ICALL has been narrowly defined as CALL software using parsers in particular. With the help of syntactic parsers, computers are able to deal with more complex tasks, such as analysing short sentences for syntactic (Heift, 1998; Schulze, 2001; Borin, 2002) correctness. As long as the domain is restricted or the range of possible errors can be anticipated they perform reasonably well. A few of these so called ICALL (intelligent CALL) applications exist Amaral (2007); Heift (1998); Nagata (1992); Schulze (2001), but only one, Compusensei (Nagata, 1992), is a commercial one.

Modern parsers are fairly accurate when it comes to dealing with language that does not contain errors. The problem is that parsers rely in some way or another on the assumption that the input they are processing is well formed and conforms to the syntactic rules of a language. Resting on this assumption that an element X has a certain syntactic function they try to establish the syntactic function of another element Y. If X is erroneous, or rather – if it does not conform to the syntactic rules – then the function of Y might not be established at all, or the computer might analyse it incorrectly, because the analysis was based on wrong premises. In order to deal with this problem various techniques have been proposed, including mal-rules (Schneider & McCoy, 1998), relaxed constraint processing (Weinberg et al., 1995; Menzel & Schröder, 1998) and learner modelling (Michaud & McCoy, 2000).

It could be argued that one of the difficulties of parsing learner language is that morphological analysers used in ICALL use Hidden Markov models (Roosmaa & Prószéky, 1998) to determine the part of speech of word forms instead of using a sophisticated parsing mechanism such as what HPSG-implementations provide. These, on the other hand, concentrate on syntax and implement only a basic morphological analysis. If a HPSG component in an ICALL application could rely on a morphological analyser to deal with the errors of a lexical or morphological nature, its task would be dramatically less difficult. It is detrimental to the state of the discipline that available morphological analysers that perform well have so far not been released under a licence that would allow CALL developers to work at integrating them with available parsing technology in order to dramatically improve the performance of ICALL applications.

3.3.6 Natural language corpora

3.3.6.1 What are corpora

QuickAssist makes use of a large German corpus in order to provide students with additional information on German words they are studying. Because of this, it seems appropriate to deal with natural language corpora in some detail here.

Corpus research (Abeillé, 2003; Garside et al., 1997; Hunston, 2002; Meyer, 2002) is a relatively young discipline in the linguistic domain. The first computerized text corpus was created in the late 1960's, but it was largely ignored or even frowned on with Chomsky's Generative Syntax then being the most influential framework. The belief in native speakers' innate competence, thus the linguists' tentative knowledge of which sentences are to be considered grammatical and which are not rendered authentic language data that corpora can provide redundant. Interest in corpus research started growing in the late seventies and early eighties. It was not only a paradigm shift, away from the generative

framework, but also the growing importance of the area of natural language processing (NLP) that led to a number of corpora being compiled for linguistic research.

By studying corpora of authentic language one hopes to gain a deeper understanding of the system of a particular language. The knowledge gained from corpus studies is used to facilitate the creation of dictionaries, grammar books, language teaching materials, speech synthesis, and speech recognition software, to name only a few applications. A lot of insights are offered by looking at the contexts in which certain idioms, words or morphemes occur, frequency lists of certain constructions can help to assess how "natural" they are compared to another one. All this can be achieved by using simple search operation on any computerized text. Ever since the World Wide Web became easily accessible by most of us, there is no shortage of texts that can be investigated with the help of the search function of a word processor or a more sophisticated application. The possibilities of this form of research, though, are limited. We can easily find all instances of the word 'the' in any given text but we cannot, at least not conveniently, determine whether this word can precede a noun, verb or preposition. We will also find it difficult, to give a German example, to compare the use of 'das' as an article, to the use of 'das' as a relative pronoun in a text.

To get a deeper understanding of the morphology and syntax of a certain language, to name a few areas of language, more information is required than just the text. If we want to find out how adjectives behave or where noun phrases can occur in a sentence we have to be able to look at all adjectives or noun phrases in a certain text. Information of this kind and a lot of other information can be added to a text corpus to facilitate research. This process is called corpus annotation and the following section will attempt to provide an overview of a number of different annotations.

The annotation that is considered most basic in corpus linguistics is referred to as Part-Of-Speech Tagging (POS Tagging), cf. above. It refers to the process of assigning

every word in a corpus a label that indicates what word class the word belongs to and information on gender, number, person, etc. where applicable. This label is referred to as tag or sometimes morphological tag.

Whereas many corpora, especially English ones, are often tagged for parts of speech, corpora that have annotations that describe the syntactic structure of the sentences included are relatively rare. These corpora are referred to as treebanks. In order to compile a treebank the output of the tagger is taken and fed into a parser. Still, most of the important treebanks rely on the human annotators that verify the analysis of parsers.

Corpora annotators usually try to implement a system of annotation which is neutral with regard to syntactic schools, there is a need to decide on whether to use phrase structure presentation or a dependency presentation of the annotated sentences.

Apart from the two major forms of corpus annotation that have just been discussed there are other forms of annotations that have been implemented for specific research needs. The first one I want to mention here is semantic tagging. It basically works by setting up a lexicon that has semantic information encoded for every lemma. The codes assigned are numbers that encode semantic content in a hierarchical order. That means that the first position assigns a semantic superclass, the following digit a subclass in that superclass, and so forth until the final subclass has been reached. For example, it would be possible to encode the word 'girl' as follows: biological-primate-human-female-young-non derogatory... This would necessitate a six level hierarchy for encoding semantic meaning. At each level a certain number would indicate that the token can be assigned exactly one of the possible semantic attribute available at that level, with each attribute having its special subclasses (it is sensible to have a subclass primate under a class biological, whereas this subclass would never be used if it was a member of the class mineral). The fact that with a five digit number it would theoretically be possible to encode 100,000 semantically different states makes this a fairly elaborate system. A semantically annotated corpus is

used primarily to look at how frequently a text uses tokens belonging to certain semantic fields. An example would be to look at the frequency of tokens marked as derogatory at some level in the hierarchy to provide empirical data to help determine whether a writer has a bias on certain issues that might otherwise be hard to discover.

Another level of annotation is the annotation of conversations. Up to now I have assumed that annotation deals with written texts. A lot of research, however, concentrates on spoken texts. These texts differ dramatically from whatever we usually consider to be a "normal" text. Spoken texts, especially free dialogues, have a number of characteristics that the annotation methods introduced so far are not able to encode. When annotating a dialogue we have to indicate for example, to state the obvious, who utters a certain sentence, word or perhaps only a sigh, cough, etc. We have to indicate where the speaker changes and whether there are interruptions, pauses or phases of simultaneous speaking. It is also important to indicate prosodic features of utterances, i.e. where speakers use stress, to emphasize something, where they speak quietly, fast, incomprehensible etc.

Finally, another level of annotation that should be mentioned is the annotation of errors. This is especially important when it comes to the computational analysis of learner language. A critical step here is to establish a typology of errors. Errors can be categorized into morphological, syntactical, semantic and orthographic errors. Each category, of course, can be further subdivided, for instance, the class of syntactic errors can be divided into word order errors and agreement errors. Here, we already see one of the important problems. Is an agreement error in German like *die schöner Frau* a syntactic or a morphological error? The more fine grained an analysis tries to be, the more difficult it will become to establish a system of annotation that annotators will use with high inter-annotator reliabilities. The establishment of a unified typology to enable researchers to compare their data across corpora and across different languages, I believe, will not be possible for quite some time.

A discussion of corpus annotation cannot avoid commenting on the technical details underlying the process of annotation. So far a number of different levels of annotation have been discussed. It has been mentioned what sort of information they add to the corpus and in some cases the tools employed have been dealt with briefly. Next, I will turn to the details of annotating a text.

If a corpus is annotated it needs to be clear what, exactly, is part of the texts that constitute the actual corpus and what are annotations of whatever kind. There are a number of conceivable ways to distinguish the "text" from the annotations, called markup when viewed from a technical perspective. Annotations could be formatted bold, in italics, a different form or in whatever other way a modern word processor can make a text look different from other elements. The problem however, is that word processors tend to be very incompatible. MS Word files created with a current version cannot be read with older versions. Needless to say that it gets even more difficult when it comes to trying to open these files on a UNIX system that is usually used for mainframe systems, systems used when huge amounts of data, such as text corpora have to be processed. Returning briefly to the pioneer days of corpus linguistics: the word processors then were of course slightly more limited in their options of changing the appearance of a text. Annotators were restricted to the use of the 255 characters that the ASCII code provides. Although there was no standard then, one can say that whatever annotation was made to a corpus was introduced by a special character that distinguished it clearly from the text. For example, the LOB corpus used an underscore after each token to separate it from its tag:

"Hello_word1 ,_punct how_word2…"

With the levels of annotation increasing it becomes necessary to indicate not only where annotation starts, but also where it ends. One method employed was to have annotations precede by the "&" sign and followed by a semi colon: "& This is annotation; this is not. Early, corpus researchers came to agree that it was desirable to agree on a

certain standard for annotating corpora. One of the main reasons was that it took and still takes a lot of work and funds to compile corpora. The only way to make up for these investments is by distributing the corpora to others. These in turn might use a number of different corpora and are accustomed to a specific annotation style or, more important, use software to extract information from the corpora that relies on the use of a certain annotation style. The very first standard that evolved was SGML (Standard General Markup Language). Two organizations have worked on standardization of corpus annotation in the past: the Text Encoding Initiative (TEI) and EAGLES. Especially since the World Wide Web seems to be the most convenient way of sharing data platform independently, these organizations are currently trying to establish a corpus annotating standard based on the Extended Markup Language (XML) developed by the W3 organization. This is an extension of SGML and is currently on the way to be the standard for data interchange. One of its advantages is that it comes with XSLT that makes it possible to access, modify and display the data stored in XML format. The latest attempt to standardize corpus annotation is XCES. It does not define what sort of annotations to use but merely provides a framework that is flexible enough to implement a tagset tailored to individual needs and still enabling the extension and transfer of the data. The developers of XCES try to promote the suggestions by Garside et al. (1997) for the annotation of corpora. These are among others:

- The original text always has to be recoverable without loss.
- The annotations have to be removable without modification of the original text.
- Annotations should be easy to understand (have mnemonic character).
- Annotations should have a hierarchical order.

Hierarchical order was mentioned earlier in connection with semantic annotation. A hierarchical order is also possible on other levels. On the part of speech level for example

the German verb ‘gingst’ can be further defined as being a lexical verb, in the past tense, second person, singular and annotated in a hierarchical order by using a tag like <VVP2S> (Verb, lexical, Past, 2nd person, Singular).

With the development of advanced annotation software and technology, with an increase of interest in corpus linguistics of researchers with various research interests, the number of annotations desired to be added to a certain corpus will grow. The more annotation a corpus contains the less readable it will become. There is already a tendency to compile multi-level corpora. The original text and different levels of annotations are stored separately and the researcher can decide which annotation he/she wants to access together with the original text. This renders corpora valuable for interdisciplinary studies, enabling researchers to concentrate on annotations specific to their subject without having to concern themselves with annotations irrelevant for their research. The corpus annotation tools (*MPI - Language Annotation Technologies*, last accessed: 18 September 2010) developed by the Max Planck Institute for Psycholinguistics in Nijmegen, the Netherlands, enable users to create as many levels of annotation as they wish. It is also possible to include audio and video tracks.

3.3.6.2 Corpora and CALL

With regard to CALL, corpora have played an important role. Heift & Schulze (2007) trace the use of corpora in CALL back to the early 1990s. They report on a number of different applications that corpora can have in language teaching.

- Corpus data can be used to test NLP based CALL applications. NLP applications, like any other computer program have to undergo extensive testing. As NLP applications are concerned with the analysis of natural language, corpora can provide a rich set of authentic test cases. In case of NLP based CALL applications, because

of the restrictions that are discussed above, the data sets used for testing necessitates a small and specialised corpus.

- Corpus data can help in the design of CALL applications. Heift & Schulze (2007) report on projects in which learner corpora were analysed in order to determine common errors that CALL software needs to address.

- Combining NLP tools and corpora into a CALL environment. This is the approach taken by QuickAssist and some other programs, where learners are directly exposed to corpus data. Details are discussed in the next chapter.

Directly exposing learners to corpus data, is a technique that is generally known as Data driven learning (DDL). It has been practised for a number of years. Corpus data is used in this approach as a learning object. The degree of exposure of learners to the corpora however has varied in individual approaches. Johns (1991), for example, created learning materials for his learners directly. Learners would then analyse the data in order to explore the semantics of a certain word, common collocations, syntactic particulars, etc. This approach was considered labour intensive by most instructors (cf. Boulton, 2010), many of whom probably had doubts about the benefits of the method in the first place. Other researchers made the corpus and a query interface available to the students directly. Depending on the underlying technology, using the tools effectively involved a steep learning curve at times.

Contrary to Boulton (2010), I do not believe that students should not be exposed to the underlying technology. Providing them with an intuitive enough interface, they are able to access NLP tools directly, work with them efficiently and do not need a middleman to filter data and turn it into a digestable form. The QuickAssist user study will show that all participants comprising various age groups and different levels of computer literacy were able to work well with the corpus interface the application provides.

3.3.7 Lexical tools

This category comprises a number of different tools that belong to the area of NLP. Electronic dictionaries, while being fairly restricted in size and quality only a few years ago, are now a standard NLP application that most of us use on a day to day bases. There is a large number of monolingual and bilingual dictionaries available both commercially from established dictionary publishers, as well as freely available ones that are quickly increasing in size and in quality. In the simplest case, a dictionary comprises a word list. In one form or another they are used by most modern text processors that use them together with a computer algorithm to implement the automatic spell checker most of us rely on.

There is also thesauri that are available in electronic form and even semantic databases like word net are now available to the general public.

In the next chapter, I will outline the development process of QuickAssist where many of the tools that were introduced above have played an important role.

Chapter 4

Development

4.1 The Design of QuickAssist

Levy (1999) distinguishes CALL design intended to be used in the real language class room and CALL design in which

> the CALL program functions as a testbed for research and is aimed at substantiating a theory, usually a theory relating to an aspect of second language acquisition. (Levy, 1999, p. 90)

QuickAssist is such a testbed. It has been developed for and has been used to test the hypothesis that making available both formal and semantic information on vocabulary items to language learners influences both vocabulary learning and the language learning process in general. But of course, I also believe that the program in its current form can be used in- and outside real language class rooms, and that both learners and instructors can benefit from the functionalities it provides.

4.2 Design principles

The development of QuickAssist was guided by the design principles for CALL applications laid out in Colpaert (2004), which have evolved into a de facto standard for CALL software development. Colpaert adapts the ADDIE approach common in industrial software development to CALL. In short, he posits an iterative workflow that includes a needs analysis(A), the design(D) and development(D) of technology to address this need, its implementation(I) and evaluation(E).

During the analysis stage, summarized in the beginning of this dissertation, the importance of an extensive vocabulary and proficiency with word formation rules was established, as was the apparent lack of teaching materials for German as a foreign language that adequately address these needs. A review of the pertinent SLA literature has shown that extensive reading and vocabulary oriented form-focus elements can provide opportunities for learners to improve in these areas. A look at the state of the art of dedicated CALL software has also led me to believe that there is an apparent lack of student centred CALL that was to be addressed during the development process. In the following sections, I will outline the design of the software and comment on the development process. The implementation and evaluation will be dealt with in the final chapters of the dissertation.

4.2.1 Open source software and reusable software components

To develop software that I would be able to release under an open software licence, was one of the most important principles during the design and development of QuickAssist. In my opinion, CALL and ICALL have suffered considerably from restrictive licencing that is often imposed on useful components that under different circumstances would lead to better performance and faster development of CALL software. It would, of course, also

make this software easily available to learners and instructors. In this section, I want to provide a short overview over the concept of open source software.

According to Stallman (2002), the founder of the Free Software Foundation, it used to be customary for software developers to exchange code in order to help each other, increasing the speed of software development and cutting development costs. In his experience, the introduction of proprietary software and the whole concept of closed source code did more harm than good for the community of computer programmers. They were not able to adapt proprietary software to their needs since the software industry reserved the right to modify code. In 1984, Stallman quit his job at the Artificial Intelligence Laboratory at the MIT and founded the Free Software Foundation in order to form a community that provided free software to its users. Users of free software are free to copy, inspect, modify and distribute the source code, in order to adapt programs to their individual needs and to improve the code base. Launching the GNU (GNU is not Unix, a pun on recursive definitions in programming) project, the community started to develop a drop-in replacement for the operations system Unix, based only on free software. It was only in the beginning of the 1990's when GNU was able to use the Linux kernel, developed by Linus Torvalds, that they were able to offer a completely free operation system with a wide range of user level applications to the public. It is the success story of GNU/Linux (often mistakenly referred to only as Linux) that has brought the phenomenon of free software to the attention of the general public.

In order to keep free software and all projects that make use of free software free, GNU introduced the General Public License (GPL), the third version of which was released in late 2007, which is also referred to as copyleft. Instead of granting authors or publishers the right of ownership, it forces them to release any piece of software that makes use of free software under the same licence again, securing that free software stays free instead of eventually becoming proprietary software itself.

The ramifications of the idea of releasing software under the GPL for developers of CALL software are as important, if not more so, than for developers in any other domain of programming. Using this kind of software, we are able to study its functionality, adapt it to our needs and use it to our own ends, as long as we do not try to distribute it as non-free software. Especially as academics who have a duty to contribute to the advance of society, this latter point should not pose any problem.

Developers facing the task of developing a new software can nowadays look around for software already available that provides a similar functionality to what they are trying to achieve, adapt it to their needs, use it and make it available to others. To provide the results of their work for free means that more and more software is becoming publicly available, existing software is steadily improved and the individual developer's programming work decreases. While writing new programs completely from scratch ceases to be a common task for developers, the domain of software engineering is becoming increasingly important.

Software engineering, however, is not usually an easy task. There is, generally, no shortage of publicly available source code that addresses various problems in many areas of NLP. However, most authors, understandably, have only their current project in mind when trying to solve a particular programming problem, giving little thought to the portability of their software. The result is a huge code base in the public domain with little or no documentation at all that would facilitate understanding of how the program works and how it can be adapted to work in another environment. In addition, there is the problem that the number of programming languages used by NLP programmers is so large that it is impossible to be proficient in more than a few of them. This really is a pity, as many of the programs and routines that are freely available, and that may be particularly good at certain tasks, may ultimately vanish into obscurity, as the programming languages they were written in go out of fashion.

4.3 Similar software

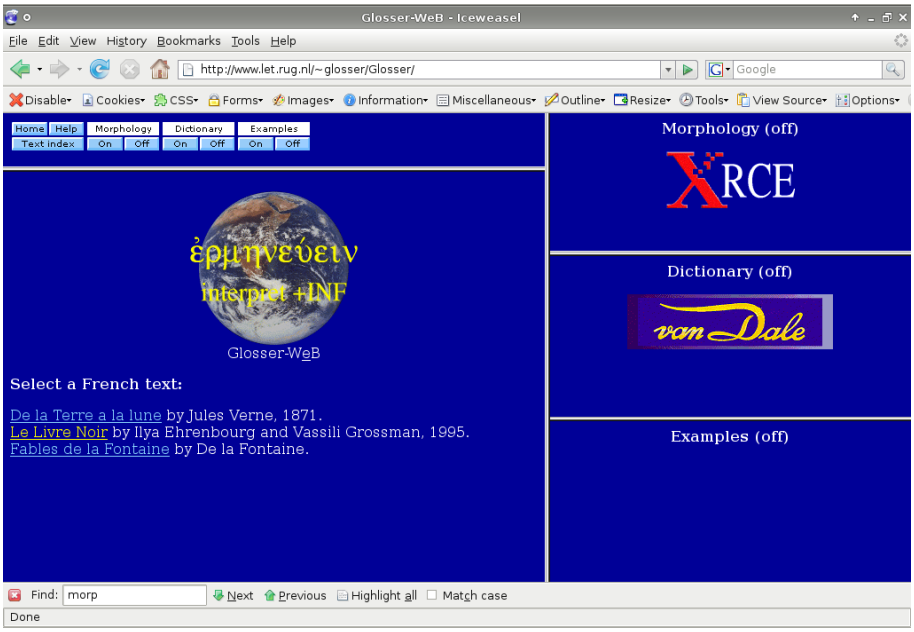


Figure 4.1: Glosser Start page

QuickAssist was inspired by software that aimed at providing learners with linguistic tools to help them read texts in the L2 or to learn L2 vocabulary. This includes concordancers and how they have been applied to data driven learning (Johns, 1991). Concordancers are usually developed for corpus linguistics. Using them is considered a non-trivial task and Johns prepared suitable concordancer output to be used by his students. Different views exist with regard to the (perceived) problem of exposing learners directly to a corpus, as was pointed out in the preceding chapter.

Glosser RuG (Dokter et al., 1998; Roosmaa & Prószéky, 1998) offered Dutch learners of French an easy to use user interface that enabled them to access concordances, morphological analyses, and dictionary definitions. While showing promising results, the project was eventually retired because of licensing issues with the dictionary that was used.

Glosser RuG was a project that attempted to develop a tool for learners that would help

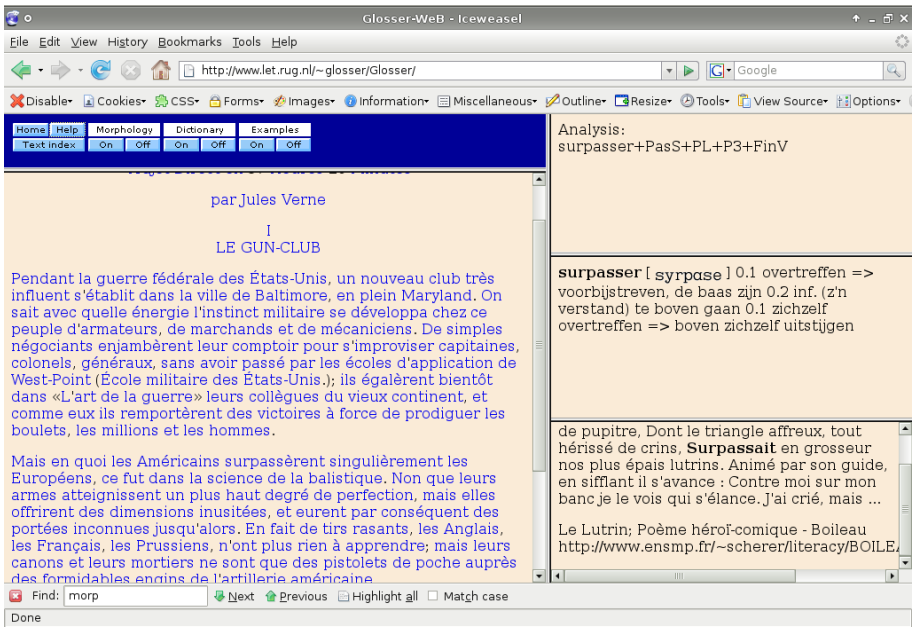


Figure 4.2: Glosser User Interface

them read a text in a foreign language by providing information on words. The resulting software is web based and makes use of the following NLP techniques: morphological analysis, Part-of-Speech (POS) disambiguation and corpus alignment.

Only a demo version used to be available for public access. At the time of writing, this public interface is no longer available. The learner is asked to pick a text from a selection of three French fictional texts from various literary periods. The text is then presented to the learner in a marked up form. The words are rendered as hyperlinks that when clicked with the mouse generate information in three different frames on the right hand side of the display. A morphological analysis of the word in question is provided in the form of a list containing the stems and abbreviations for inflectional affixes. In the frame below, the learner is presented with a conventional dictionary definition of the word in Dutch and the bottom frame displays a sample sentence retrieved from a French corpus that shows the word in a different context. Each of these frames can be individually toggled on and off.

This gives learners the opportunity to work through the text with varying degrees of help from the computer. They can try to infer the meaning of an unknown word from corpus examples if the original context is not helpful or ambiguous, can use the morphological analysis to help them infer the meaning, or simply look up the dictionary definition.

This software enables the student to access a natural language corpus and actively research the meaning of an unknown word. Therefore, it is much more valuable than a conventional dictionary that is only able to provide a narrow definition of the word and will often fail to account for all usages of the word in question. Also the dictionary look-up falls short of providing a high degree of the learner's active involvement in the meaning inference process. Chapelle (1998) emphasises that a computer can play an important role in facilitating noticing. This is the phase in an Input-Output-Interaction model of Second Language Acquisition (Gass & Selinker, 2008) at which the learner notices a new language feature, the crucial initiation of uptake and finally acquisition. QuickAssist provides the learner with a keyword-in-context option, optionally displays the inflectional paradigm, and highlights forms in the information pane. This is intended to help learners not only to notice new phenomena, but to guide them from there at a student determined pace, with optional rule explanations and the study of corpus data along all of the stages of the acquisition process.

Cyberbuch (Chun & Plass, 1996) mainly used multimedia annotations like images, videos, and audio files to increase the saliency of vocabulary items. Since these can also be considered glosses it is certainly related to Glosser and QuickAssist. Cyberbuch has been used intensively over a number of years to study vocabulary acquisition in normal online learning. In this case it functions like a testbed for SLA theories in much the same sense that QuickAssist does. Cyberbuch comes with a set of texts that users can work with. The glosses that are made available to users had to be created manually by the software developers. Although both the texts and glosses contained in Cyberbuch were

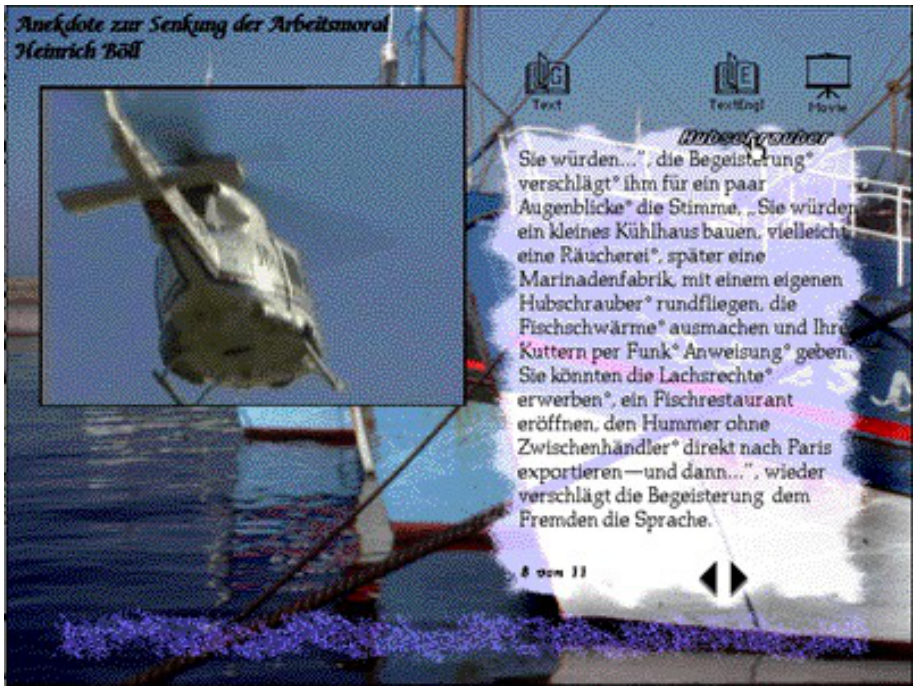


Figure 4.3: Cyberbuch

designed with a particular audience in mind and address their specific needs, it is not as flexible as Glosser RuG and QuickAssist with regard to the freedom of choice users have in respect to the texts they can work with.

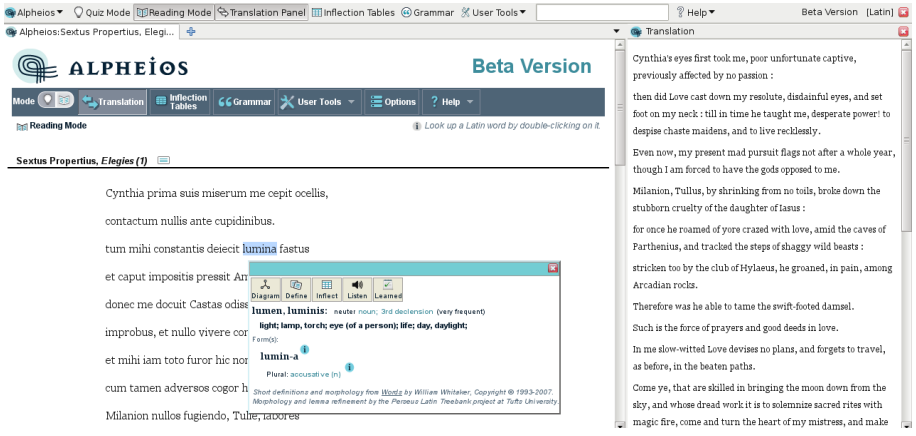


Figure 4.4: Alpheios

Only recently, Alpheios (*Alpheios*, last accessed: 17 September 2010) introduced an online glossing tool for ancient Greek and Latin which uses a Firefox-plugin as a user interface. Users are required to install two to three plugins for Firefox, depending on whether they want to be able to work with both Greek and Latin texts or only one of them. In the currently available beta version, users can select a text and read it with the help of some tools:

- An English translation of the text that has been aligned with the help of corpus alignment techniques can be displayed in a side pane, or toggled off. If the pane is displayed, moving the mouse over a word in either pane, will display the position of the equivalent in the other language in the other pane
- Users can also check the translation of a word by double clicking it
- An annotated syntax tree of the sentence they are currently reading can be displayed to users
- A media player can be used to play back individual words (this function did not work when I tested the application on my system)
- Users can look up grammar explanations in a grammar book
- Inflectional paradigms of nouns, verbs, and adjectives can be displayed
- The quiz mode offers users exercises to practice vocabulary items that they have selected while reading the text

In terms of functionality as will become clear in the following chapter, Alpheios is fairly similar to QuickAssist. Offering only limited texts, however, it is able to provide additional functionalities, that QuickAssist cannot provide, such as the corpus aligned translations.

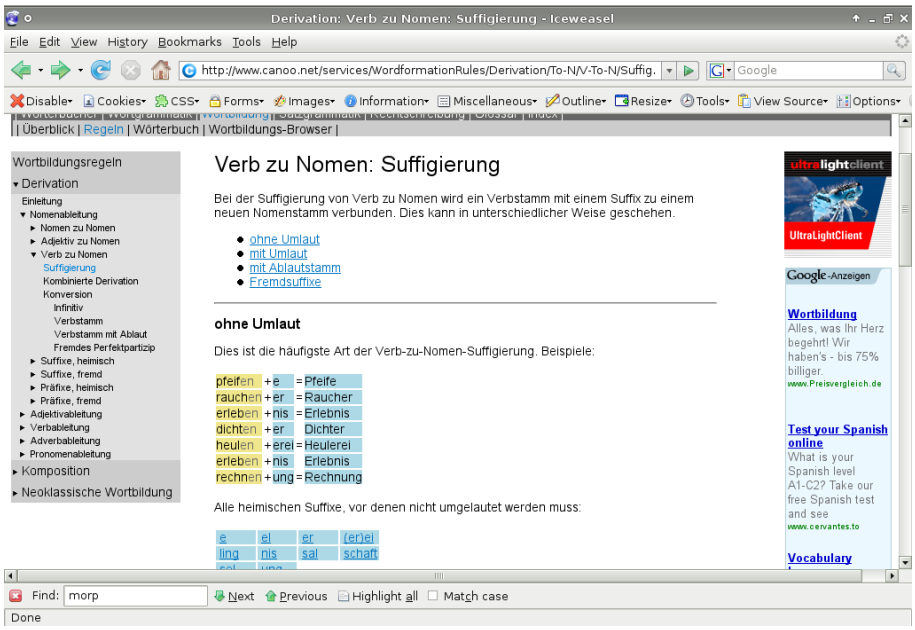


Figure 4.5: Word Manager: homepage

Other applications that are not strictly CALL applications in the narrow sense have influenced the development of QuickAssist:

Word Manager (Hacken & Domenig, 1996) was developed by Marc Domenig and others. It is a morphological database which is designed to be accessed by various NLP tools and to provide morphological information on word forms such as the lexeme it belongs to, inflectional and derivational morphemes it contains and information on compounding if applicable. The database interacts with other NLP applications in a client server architecture were the database provides different interfaces to different users that can be used to submit queries, retrieve results and to modify the database in some cases. The database is composed of a rule part that is available for modification and expansion via the linguist interface and the part holding morphemes and lexemes that can be accessed through the lexicographer interface. The use of these two interfaces are restricted to developers. The public interface can not modify the database and is used to communicate with other

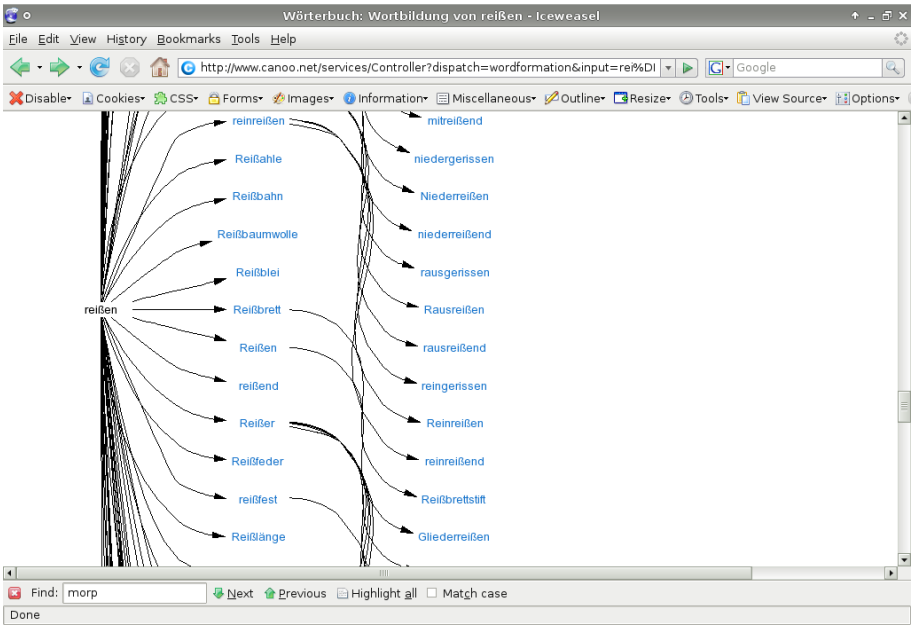


Figure 4.6: Word Manager: display of related words

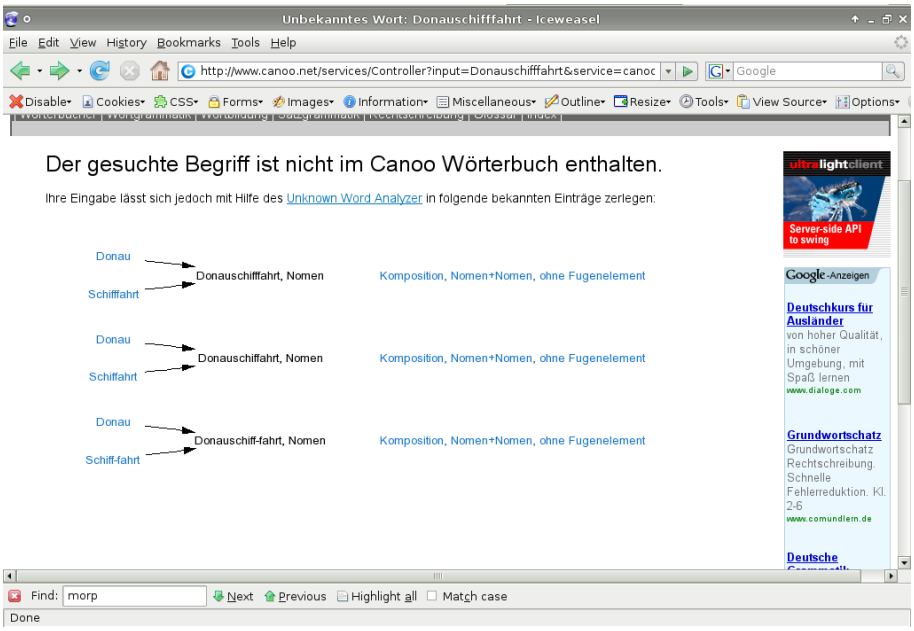


Figure 4.7: Word Manager: morphological analysis of words that are not listed in the dictionary

applications.

Word Manager as such is proprietary software and not freely available. There are however internet platforms available to language learners that are based on the Word Manager system. The one I will concentrate on is for German and available at http://www.canoo.net, but other systems like the German Italian ELDIT (Hacken et al., 2006) exist or are currently being developed.

Canoo.net offers the user a wide variety of options. Different dictionaries can be consulted to look up a word in various languages, information on pre and post spelling-reform variants is available, there is a short online grammar with information on German morphology and syntax and words can be morphologically analysed online. The morphological analysis is presented to the user in the form of a tree diagram which shows how compounding, derivation, and inflection were applied to derive a particular word form. Diagrams of the word form in relation to other members of the word family it belongs to can be generated as well as inflectional paradigms of nouns, verbs, and adjectives.

While there is a plethora of options available and detailed morphological analyses can be created, there is no interface available to enable other NLP applications to access the information that Canoo.net offers to the human user. It is possible to create search forms that link to the website and to download add-ons for MS Word which enable the user to check words directly from within the word processor. The missing public Application Programming Interface (API), however, is a major drawback since it ultimately makes it impossible to create CALL applications for the public domain using Word Manager or any system based on it.

For QuickAssist, I have chosen to provide users with the ability to look up the morphological analysis of individual words and to display their inflectional paradigms. Given the wealth of information and sub menus on the site, this is intended to help learners find useful information fast. This decision was based on the observation that many of my

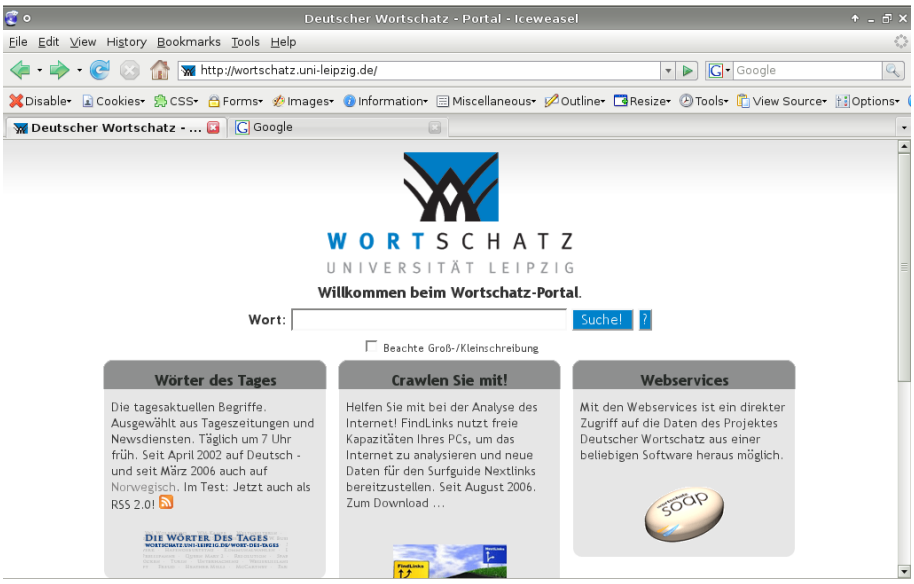


Figure 4.8: Wortschatz: homepage

students seemed to have problems making efficient use of the web site.

Wortschatz (Quasthoff et al., 2006) was developed at the University of Leipzig in Germany. The user is offered to look up a word form with the web interface and is provided not only with a morphological analysis of the word form but can also access information on its part of speech, typical collocations, as well as semantically related words and semantic fields the word belongs to. It is also possible to access a KWIC list that is generated with a corpus look-up. The database is also accessible through a SOAP interface (Curbera et al., 2002) that enables NLP programs to send queries to the database and receive the results in XML-format (Bray et al., 2000). This facilitates the development of modules that can interact with the Wortschatz database and exploit the information provided in a CALL application. The project also makes the database available for download to researchers.

This section showed that the idea of using computers to help in the acquisition of

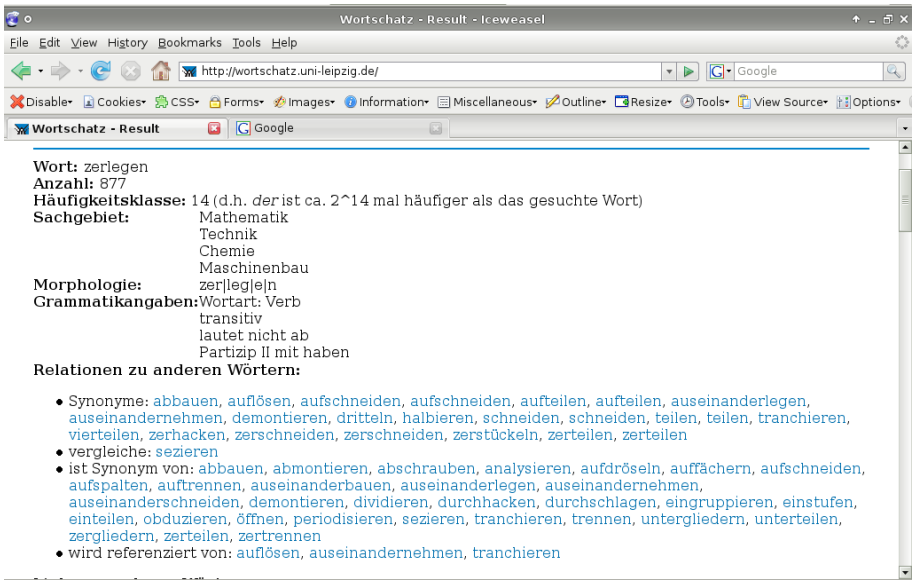


Figure 4.9: Wortschatz: information on a word

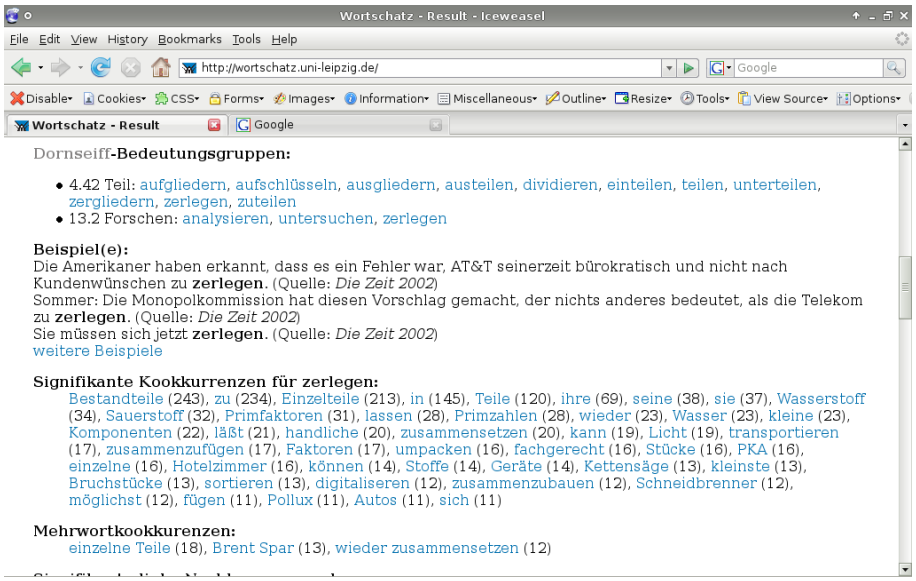


Figure 4.10: Wortschatz: corpus look-up and information on co-occurrences of a word

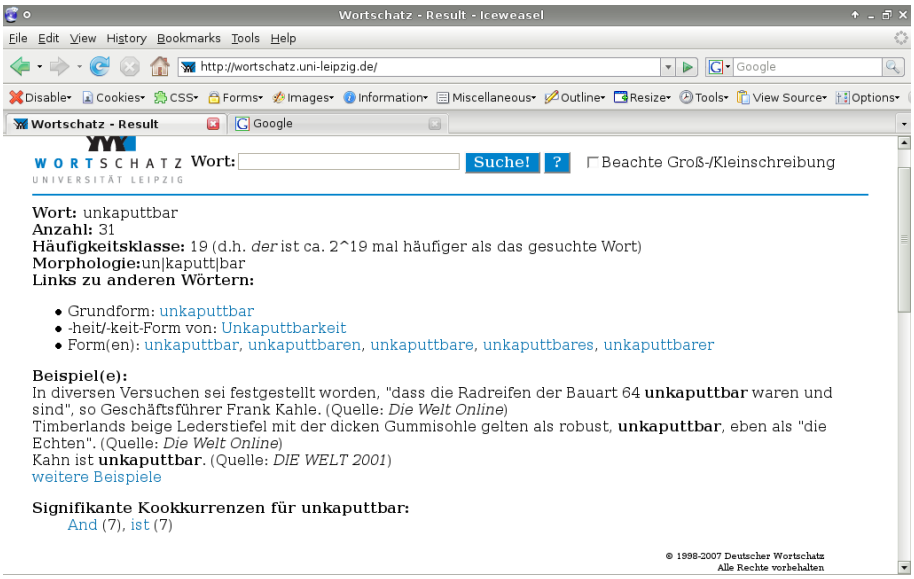


Figure 4.11: Wortschatz also provides some information on words with unconventional morphology

vocabulary and word formation rules is not a novel one by any means. This overview shows that there are a number of possibilities that NLP techniques offer and it explains how they are exploited by other projects. It does not seem feasible nor desirable to develop a completely new system hoping to achieve better results than the others especially given the restraints on time and resources that come with this project.

There is, however, room for improvement: While Glosser seems to be an interesting learning aid that can be used by learners of all language levels and with varying degrees of independence, the tool is only available for Dutch learners of French. Word Manager is proprietary software and cannot be used without incurring high licencing fees, and Wortschatz, while providing corpus look ups and a lot of other features is, by itself, of limited use to all but the most advanced language learners. As demonstrated in Wood (2007) there are some problems with the morphological analysis of some word forms both in Word Manager and in Wortschatz which could be improved.

The most promising results can be achieved by borrowing ideas freely from other projects and create similar features that QuickAssist provides to the learner, and use the publicly available API that Wortschatz offers.

4.4 Finding a suitable programming language

In retrospect, finding a suitable programming language for the project was probably one of the most time intensive stages. This is largely due to the fact that I consulted with a number of people and read a lot of the pertinent discussions in literature. There is a vast number of programming languages and there is a good reason for this variety. Programming languages usually evolve out of the need to deal with certain tasks in the most efficient way.

Especially when it comes to the domain of Natural Language Processing, there are a number of programming languages, from different families, that have traditionally been used to deal with certain tasks. In the area of syntactic parsing and morphological analysis it used to be Prolog and Lisp that were widely used in many projects. This popularity is due to the nature of these languages. In contrast to so called procedural languages for which the order of commands determines how the program is executed, these languages are based on rules. Programming in these languages means defining rules and evaluating expressions that make use of these rules. Given that rewrite rules such as the ones used in transformational syntax models can be easily implemented in a logic programming language like Prolog made them popular with computational linguists and they are still used to some extent, especially in educational contexts.

It is probably a matter of speculation why Prolog (Gazdar & Mellish, 1989) ceased to be widely used, but a few of the reasons for it might have been performance issues, competing standards and the steep learning curve involved when trying to learn Prolog

coming from a background in procedural languages. More recently functional programming languages like Haskell (Hutton, 2007) have been used in some NLP projects. While they are somewhat similar to logic programming languages in that they are rule based and work largely with recursive algorithms, etc., there are certainly no more serious performance issues associated with these languages. It was mainly because of the large amount of time that I would have spent to become sufficiently proficient in Haskell to be able to work productively on the project that I looked for an alternative.

C (Kernighan & Ritchie, 1988) and it's successor, C++ (Stroustrup, 1997) have been used in the Unix world ever since Unix (Bach, 1986) itself was programmed in C. Both languages are praised for their performance as well as for the availability of libraries, modules, and toolkits to create Graphical User Interfaces (GUIs) on a wide variety of different architectures. I spent quite some time learning these languages and also started developing early prototypes of QuickAssist in C++ using QT (Blanchette & Summerfield, 2008) as a toolkit to develop applications that could run at least on the three major operation systems, Windows, Macintosh and Linux. The problem with both languages is that development is quite time intensive because both languages are low level languages, require a lot of code, compiling and debugging to implement relatively common functions.

Especially to achieve faster development times and to avoid time intensive debugging, I started looking for alternatives and experimented with Perl (Wall & Loukides, 2000) and Python (Lutz, 2006), scripting languages, whose main advantage is that they require relatively little code, are easy to debug and allow for rapid application development. While both languages have a relatively similar syntax which is in fact based largely on C/C++, I decided in the end to go with Python, because code in Python tends to be a lot more structured than Perl code. An early implementation of QuickAssist was presented at the CALICO conference 2007 and also at the ACLA-CAAL conference in the same year.

It was during one of these conferences that I had a long conversation with a computer

scientist working in the industry who convinced me to give Java (Arnold et al., 1997) a try. His argument was then that Python might not be able to scale well once a fully fledged corpus and dictionary was available. I still think that Python would have been able to perform quite adequately, but there is an even stronger argument for Java. While early versions of Java were shunned by many developers who claimed that its performance was quite poor, this has not been a major issue for some time. Java is available for all major operation systems, is easy to install and while Java applets where frequently used in web development, the vast majority of computer users had it installed on their system. Java is still used heavily for the development of applications that run on web servers (Hunter & Crawford, 2001). It never got very popular as a development tool for applications intended to run on web clients and with Adobe's Flash (Hall & Wan, 2002) becoming ever more popular, there is not much reason to believe that this trend will be reversed.

The main reason for this is very likely that Java applications depend on a so called virtual machine which takes precompiled Java byte code and executes it. Because Java is a language that has a very extensive set of libraries, this virtual machine is relatively resource hungry, compared to Adobe's Flashplayer. When it comes to desktop applications, however, this is no real issue. The virtual machine can be distributed with the application. In the case of QuickAssist the roughly 130 megabytes that the installation software for the virtual machine requires dwarfs in comparison with the more than three gigabytes that the databases require on the DVD.

In terms of cross platform compatibility, Java is fairly hard to beat. While it could be argued that Perl or Python are equally platform independent, the fact remains that in order to achieve true platform independence with these languages, some things cannot be taken for granted, such as the character encoding. While many users of a Windows system will use an ISO-8859-1, or an ISO-8859-15 encoding, other users might use UTF-8 encoding by default. Taking into consideration that the software might be used by learners who

have their standard encoding set to a non Western encoding, programming routines that do text processing of any sort can get very complicated. With Java, this problem does not exist because it works with a sixteen bit UTF encoding internally, no matter on what platform it runs. In addition, it is probably easier to download and install a Java runtime environment than it is to set up Perl or Python on a system where these languages are not installed by default.

4.5 Finding suitable components

In this section, the components that QuickAssist uses are described. First, the Java technology used in QuickAssist is described.

Based on this discussion, it is possible to motivate how the actual NLP components that QuickAssist offers to the user are integrated in the application. The corpus and word-form list that lie at the center on QuickAssist are discussed in some detail.

4.5.1 Java Components

4.5.1.1 The Standard Widget Toolkit (SWT)

Java comes with its own libraries that support GUI programming, called SWING libraries. These libraries are very extensive and perform quite satisfactory on most computers. Since design issues were fixed in recent versions, the windows created with SWING even look like the other windows on the user's desktop.

Early on in the development process it became clear that it would be of advantage if QuickAssist could behave quite similar to a web browser. The rationale behind this was that there would be hardly any learning curve for learners, who, as long as they had some proficiency with the use of web browsers, would find it easy to work with QuickAssist as

it behaves in a very similar way. That this is indeed the case was established in the user study. We will return to this point in the pertinent chapter.

SWING offers widgets (GUI elements) that function similar to a browser in so far as they are able to render very basic HTML. When experimenting with first prototypes it became clear, though, that the functionality of these widgets was far too restricted. They could not interpret cascading style sheets (CSS) and it was thus not possible to determine the layout of the text. Since every word in QuickAssist has to be "clickable" in terms of HTML this means that every word has to be defined as a link. SWING can work with links, but renders them in a standard way so that they are underlined and once the user clicks on a link, it changes its colour. Both of these behaviours were undesired and rendered the text in the window hard to read.

It is good practise in software development to restrict the number of third party libraries to a minimum. Not only do the resulting applications get bloated if too many libraries are included, developers have to learn a new application programming interface (API) for each of theses libraries. In addition, maintenance of the software gets more complicated, because it is necessary to make sure that the individual libraries are still compatible with each other when one or more are updated. If one considers all of these problems, there has to be a good reason to include a third party library, especially if Java already seems to provide the functionality that a new package offers. In the case of SWT, the advantages outweighed the conceivable disadvantages. SWT was developed for the Eclipse project and is still actively developed largely with the help of IBM. The GUI uses routines of the underlying platform, which on most platforms leads to a better performance and more authentic look and feel of the user interface than is the case with SWING applications. But most importantly, the browser widget that SWT uses does not imitate a browser the way SWING does – it integrates the system's standard web browser. On a Windows system Internet Explorer will be used, Safari on a Macintosh and usually

Firefox on a Linux machine. That means that the full range of functionalities that one can expect of a modern browser is available.

Not only did SWT solve the problem with the links that was outlined above, it made it possible to use authentic webcontent in the browser windows. In the case of QuickAssist this means that it is possible to load for example a Wikipedia page and let the user interact with it in the same way she would with a normal webbrowser. She is able to read a webpage, scroll through it and follow links on the page. Providing a similar functionality using only SWING components would have necessitated far more development time than what was available

4.5.1.2 The Derby Database

The second and last external library that is used for QuickAssist is the derby database. Derby is not a third party library in the strict sense as it used to belong to Sun. It has meanwhile been open sourced and is maintained by Apache. It was originally planned to make it part of the Sun Java 1.6 distribution, but this plan was not realised. Whether this may happen with a later version of Java is a matter of speculation, but given that Sun acquired MySQL some time ago and was recently taken over itself by Oracle who will most likely be most interested in promoting their own database, this is very unlikely.

It would have probably been the easiest to design QuickAssist as a client application that connects to a remote database in order to retrieve data such as dictionary entries and corpus data. Especially the distribution of the software would have been facilitated that way, because the program itself and the required libraries would have required under six megabytes of disc space and could have been made available for download from a website or even sent as an e-mail attachment.

The size of the database that would have had to be installed on one of the university's servers, the time and effort that this would have had required on the part of our IT de-

partment was felt to be unproportionally high for this research project. It was decided to avoid administrative problems by using a local database, and to distribute the application together with the database on a DVD-ROM. This might not be an ideal scenario if a wider audience is targeted, but proved to be a feasible method to make the application available to the participants of the user study and anybody else interested in the software. In addition, this way users are not dependent on a fast internet connection and do not have to fear that some server might be temporarily down. Most of QuickAssist's functions are available off-line.

Derby was used for a variety of reasons. First of all, it is a very small application. It does not require any complicated installation procedures, and it does not even require a database server to be running. Quick Assist connects to Derby in embedded mode which allows only single connections, but results in a very small memory footprint. In addition, with Derby, the entire database, as well as the application itself can exist in read only memory so that QuickAssist can run entirely from DVD, no disk space on the users hard drive is required apart from a few temporary files that are created and deleted while the program is running.

4.5.2 NLP Components

4.5.2.1 Description of the Corpus

As was discussed earlier, Corpora have been widely used in second language teaching contexts. Usually, however, the use of corpora in data driven learning involves the use of a concordancer, which is a useful NLP application, but is complex and using it involves a steep learning curve for the learners or their instructors who only make the output of the concordancer available to their students (Johns, 1991). QuickAssist offers an easier and potentially more effective approach. It integrates the corpus as a resource in a ded-

icated language learning tool and offers users an intuitive interface hiding much of the complexity.

It was not a question whether QuickAssist should use a corpus, but rather which one to use. This question, however, proved difficult to answer. There were two requirements that a corpus had to meet to be suitable for the project. The first requirement was that it had to be a freely available corpus of contemporary standard German, otherwise the software could not be distributed under the GPL, version 3.0. The second requirement was that the language stored in the corpus should be fairly easy to understand for intermediate to advanced learners of German.

The original idea, to create a custom corpus for the project was soon abandoned. It would have certainly been beneficial to have full control over the contents of the corpus, especially since this would have allowed to target a broader audience by providing corpora suitable for learners with different levels of proficiency. Creating a corpus, however, is not a trivial endeavour. It would have involved finding suitable texts from a variety of different sources in order to create well balanced content. It would have been necessary to find out for each of these texts whether it was protected under some copyright law and if that would have been the case, soliciting the permission to use the text would have been necessary. In addition, the corpus would have had to be cleaned, which is another labour-intensive process.

There are a number of German corpora available. Quite a few of them can only be used if a licensing fee is paid. This was of course not feasible as there was no funding available for such licensing fees, nor was it desirable that potential users of QuickAssist would have had to pay a licensing fee for third party products. Other corpora offer a web interface that is open to public use. People can use the web interface to create queries and receive the results of it on their browser. Using such a corpus would have had the disadvantage that QuickAssist users would have had to be redirected to that web interface any time

they wanted to access corpus information, or that a routine would have had to be written to generate the request automatically and then parse the resulting html file that would be generated as a response. The first option would have been less than ideal, because instead of accessing corpus information through the click of one button, users would have had to go through the cumbersome process of creating the queries themselves. Taking the second route would have meant that any time the web interface changes, which frequently happens because of updates to the underlying software or merely because a new design is implemented, the routines to generate a query automatically and to parse the results would have required modification.

The corpus that was chosen in the end was the 300.000 sentence corpus from the Wortschatz project at the University of Leipzig (Quasthoff et al., 2006). Wortschatz makes the various corpora which are part of the project available in several forms. The German corpora mainly comprise newspaper articles. The range of resources used is impressive:

> 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.
>
> 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 et al., 2004).
>
> Source: Pokorny (2009)

Like many other corpora, the Wortschatz corpora can be queried through a web interface, but it is also possible to query the corpora through a SOAP API, a standard for web based resource access. This standard guarantees that routines accessing the resource through the API will always work, because the API will remain the same even if the underlying functionality changes. Early Python versions of QuickAssist used the SOAP interface to access the German Wortschatz corpus. Internally the Wortschatz corpora are implemented as a MySQL database, a very fast and freely available relational database which is widely used in industry and research. Wortschatz makes these databases available for download as long as they are used for research purposes. This range of different access methods certainly makes Wortschatz a unique resource.

Because initial experiments showed that access through the SOAP interface was slow sometimes, that the server regularly went down for maintenance, and because QuickAssist needed its own database anyway for the dictionary, the full form list, etc., I decided to use the MySQL database files just mentioned. As was explained earlier, QuickAssist uses Derby, not MySQL as a database application, mainly because it is far easier to install than MySQL. This made it necessary to convert the MySQL database files into a format that Derby can use.

The database files from the Wortschatz project were downloaded and installed on a MySQL server. The tables were then individually dumped to plain text files, a process that is normally used to backup a database to safeguard against information loss. The resulting text files were SQL (Standard Query Language) scripts that, fed to a MySQL server, would recreate the tables. Both Derby and MySQL use SQL, but while Derby adheres closer to standard SQL, MySQL uses a range of commands and more importantly data types that are not used by other vendors. To deal with these idiosyncrasies, I developed some Python scripts to change data types that are not part of Derby's SQL syntax to types that Derby can process. After these scripts had been executed, the resulting SQL batch

files could be processed by Derby to create Derby tables. In order to achieve fast retrieval rates for the different queries that QuickAssist performs, a number of indices were also generated.

As the structure of the Derby tables are identical to the Wortschatz tables, a detailed description is not necessary here. A wealth of information can be found in Quasthoff et al. (2006) and in the excellent documentation that is available on the Wortschatz site (*Leipzig Corpora*, last accessed: 13 September 2010). In short: we are using the 300.000 sentence corpus in QuickAssist. While bigger corpora are available, they would have been too big to distribute on a DVD and using a bigger corpus would most likely have only increased the amount of available sentences, not so much the range of vocabulary covered, since the corpora are all compiled using the same sources which are newspaper sentences. The corpus was cleaned using heuristic methods to assure that only German sentences are included, that lists, and sentence fragments are excluded and (this is not mentioned anywhere in the Wortschatz documentation) that sentences have a maximum length of 255 characters. The database contains a table containing all the sentences and a number uniquely identifying each of them, a table containing the individual words and a unique number for each of them. There is one table that is used to find out what sentences a certain word occurs in. It contains a row for every word listing its index number, the index of the corresponding sentence and the frequency of the word in the corpus. In addition there is one table that is used to look up information on the co-occurrence or words, by listing the indices of both words, the frequency of their co-occurrence and a significance measure.

QuickAssist uses these tables to generate keyword-in-context (KWIC) lists, to find direct neighbours of a word and to retrieve information on the frequency of a word the user selects. These functions were included to provide the user with an array of possibilities to independently infer the meaning of a word, learn about possible contexts, lean more about

its semantics, possible other meanings and how commonly used it is in contemporary standard German, at least in the domain of newspaper articles.

4.5.2.2 Description of the Wordform list

In order to find the lemma for a certain word form, an NLP application can take several routes. For some languages it might be sufficient to use a stemmer. This application checks forms it derives by chopping of characters from the end or from the beginning of the word against a list of stems. If the form is found in a list, it becomes a possible stem. This works quite well for languages like English were the plural for most nouns is formed by appending s to the end of the singular form. While there are stemmers for German, though, their performance is not good enough if they are expected to produce accurate results for individual words rather than predicting the stems of words reasonably well to be used successfully in statistical language processing contexts.

The best results would be achieved with a morphological analyser. These programs have access to a wealth of morphological information and can be used to identify the morphological constituents of a word, rather than stems usually using a series of finite state automata. While morphological analysers would be the most effective tool to be used for QuickAssist, it was not possible to do so. The only working morphological analysers I found for German are proprietary software. The plan to develop an open source morphological analyser as part of this project was soon abandoned, because it would have taken far too long to implement it and get it to work well enough to produce comparable results to the available products. While canoo.net provides free access to Word Manager, a commercial tool for morphological analysis, it does so only via a web interface and the same problems that exist with respect to German corpora which only provide a web interface for public use apply to canoo.net. In this case, as will be outlined later, a solution was found to enable learners to benefit from the content offered on canoo.

Having established that stemmers were not able to provide satisfactory results, that a morphologic analyser was not available, the search for alternative solutions led me back to a project I worked on in 2005. I developed a tool for the semi-automatic analysis of learner language for the WatPal project lead by Mathias Schulze. As part of this project I developed a Perl script that enabled us to automatically calculate the type-token ratio of a text. This of course made it necessary to find the appropriate type/lemma for every token/word in a text. The way we did this back then was to use a word list that contained a wordform and the corresponding lemma on each line. In order to get this list, we used Morphy, a program that was developed by Wolfgang Lezius (Lezius, 2000). While the source code for the program containing the routines that are responsible for the morphologic analysis of words is closed source, the program itself is available for free download *Morphy* (last accessed: 17 September 2010). It is possible to create a list of all the wordforms that Morphy knows about and run this list through the morphologic analysis filtering out all the information apart from the Part of Speech (POS) information and the information on lemmata. We used the resulting full form list as a plain text file in the Watpal project and scanned the file for every wordform in the texts that had to be analysed. This is arguably a very inefficient way of information retrieval, but as the analysis did not have to be performed online and time was not an issue, this was the easiest way of implementing the required functionality and did not make use of any complex packages such as a database. In order to use the full form list efficiently in QuickAssist, that is to look up the lemma for any wordform without any delay that would be noticeable to the software user, the lookup process had to become far quicker. This is why the fullform list had to be converted into a database table for the Derby database. A Python script was used again to create an SQL batch file that, passed to Derby, created the table and build the search index to help the database perform speed efficient lookups.

4.5.2.3 Other NLP components

In order to be able to provide a sophisticated morphological analysis of German words and other functions it was important to find a way of interfacing with *Word Manger* (Hacken & Domenig, 1996). This is by far the best tool of its kind for German. Its functionality is available online on the Canoo web site (http://www.canoo.net). Regrettably, Canoo.net, unlike Wortschatz, does not offer an API for applications to query the database directly. In order to make some of the information that can be accessed on the website easily available for users, it was decided to call the website with a pre-set query in the URL address. This way, users are able to request a morphological analysis or an inflectional paradigm with the click of one button instead of working their way through Canoo.net's comprehensive set of menus and sub-menus.

While it would have been better to implement the functionality that Word Manager offers for various reasons, such as being not dependent on the canoo.net service (which was reportedly not available on some occasions during the user study), offering the user a unified interface, and the ability to customize the output, this was simply not possible in a project limited by constraints on time, financial resources and man power.

It was fairly easy to find a bilingual German English dictionary that I was able to use: *Freedict* (last accessed: 13 September 2010). This, as well as the thesaurus, that is used to find synonyms (*Open Office Thesaurus*, last accessed: 13 September 2010) does not offer all the features commercially available packages might do, but both have a surprisingly broad coverage. Moreover, since they consist of simple lists, it was relatively easy to adapt them to be used in *QuickAssist*.

In the early testing stages, QuickAssist also interfaced with the Stanford Tagger (*Stanford Tagger*, last accessed: 13 September 2010). This made it possible to identify most of the separable verbs in texts used for testing. As the tagger has a fairly large memory

footprint and can require users to modify the configuration of the Java virtual machine, it was decided to not use it for the user study in order to keep technical problems that users potentially had to face when installing and using the application on their own computers to a minimum.

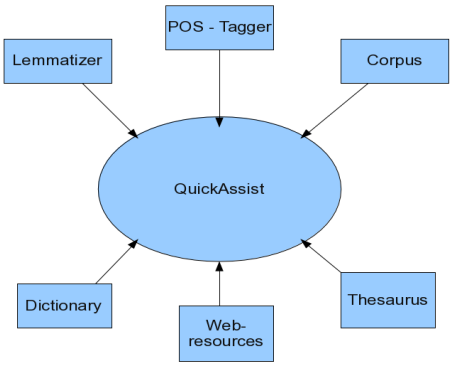
4.6 Architecture

Figure 4.12 gives an overview of how the individual modules interact. Most of the resources are accessed through a query to the QuickAssist database. It contains the list of fullforms generated by Morphy, the German corpus obtained from the Leipzig Corpora Collection, the dictionary and the thesaurus. In addition, the application offers access to the Canoo.net website for morphological analyses and other useful information and to the German Wikipedia. This is especially useful to find information that has not yet made its way into dictionaries. When reading newspaper articles and similar texts, users are able to quickly access information on persons, places and events, and might even find a helpful image. QuickAssist can also be used offline and will still provide all the functionality excluding the web resources.

The following list illustrates how the application handles user requests that are triggered by pushing one of the function buttons:

- User presses the *Englisch* button
- Application determines the currently selected word form
- It sends a database query to derby to determine whether the wordform has a corresponding entry in the wordform–lemma list
- If the request returns a corresponding lemma, the application queries the database for an English translation for the lemma

Figure 4.12: Architecture of QuickAssist



- If the request returns nothing, the wordform itself is queried in the database
- If the query for a translation returns results these are displayed with the search term highlighted
- Otherwise a message is displayed that no suitable entry was found

To be able to develop an application that can run on different computer systems, I decided to use Java as a programming language. QuickAssist was tested on different Linux distributions and on machines running different Windows versions. It should also work on Macintosh computers, as well as BSD and Solaris systems. In order to be able to use sophisticated web browser features that exceed the capabilities of Sun's graphical user interface (GUI) kit Swing, QuickAssist makes use of the Standard Widget Toolkit (SWT) that is developed by the Eclipse project (*SWT*, last accessed: 13 September 2010). It embeds the system's web browser, so it is necessary to install a suitable version of SWT in QuickAssist's directory. Since this enables accessing both the Canoo website for

morphological analysis and the German version of the Wikipedia, this is an acceptable trade-off. Most users will also find that the SWT GUI looks more “natural” than Swing and easier to use.

From a human-computer-interaction (HCI) perspective, QuickAssist should be easy to use and intuitive for most users, as the user only has to deal with a few function buttons and will soon find that the application behaves almost like a web browser. The database used is Apache Derby. Its advantages are that it is a Java application which makes it platform independent, it has a small footprint and can even be distributed on a DVD ROM. This makes it possible to work without an external database server. All the data is distributed together with the application on a DVD. Derby is a relational database that can be queried with SQL. Although it is somewhat different from MySQL it was relatively easy to convert the Leipzig corpora that are available in MySQL format to a format that is compatible with Derby.

Chapter 5

Implementation

This chapter describes briefly the functionalities of QuickAssist.

Figure 5.1: QuickAssist Startup

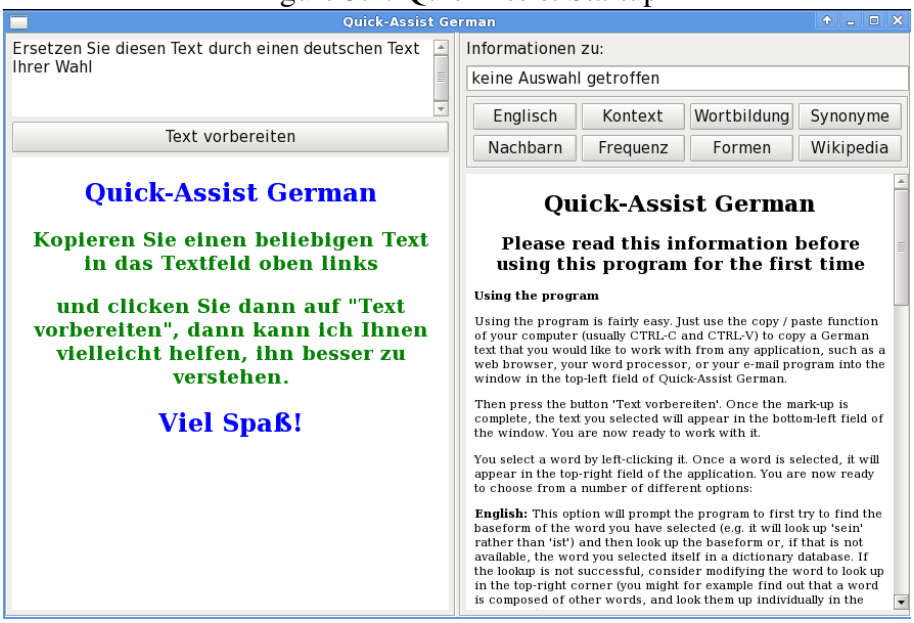


Figure 5.1 shows the start-up screen that the user is presented with when launching the application. The number of buttons, menus, etc. is kept to a minimum. The pane in the top left of the window is the region where users paste a text they intend to work

with. This, as the other panes, on start-up, displays a helpful message. The button below triggers the mark up process. The text is turned into a hypertext. Each individual token is turned into a hyperlink that can be clicked. The marked-up text appears in the bottom left pane. When the user clicks on a token in the marked up text, the token will be displayed in the input area at the top right.

Below this input area are the buttons triggering QuickAssist's individual functions. The output that these functions create will be displayed in the bottom right panel. In order to enable users to read the output comfortably, the divider between left and right elements can be moved. This is useful to extend the width of the right-hand output pane when displaying a complex website like the Wikipedia, or Canoo.net.

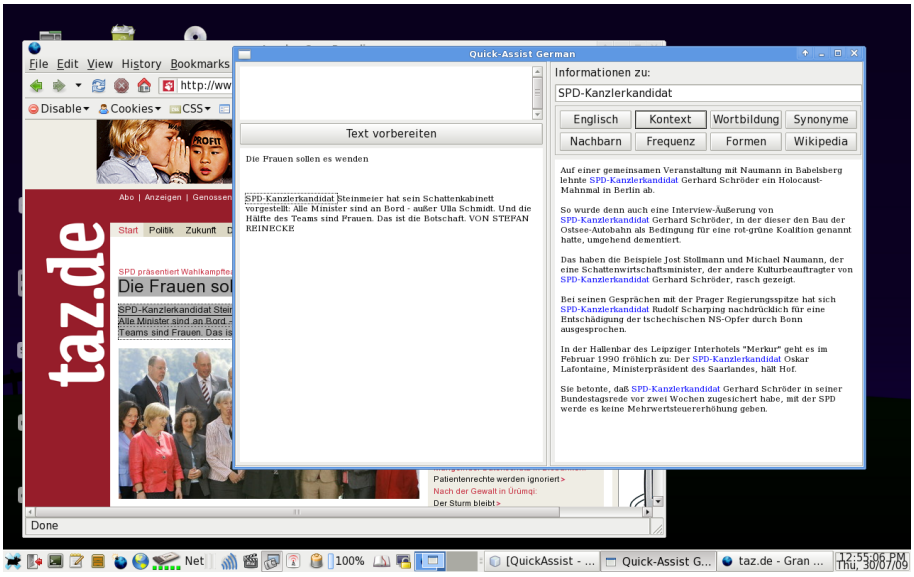


Figure 5.2: QuickAssist: KWIC View

Figure 5.2 shows the basic workflow of using QuickAssist: users select a text they want to work with. Usually this is done using internet resources, but it is also possible to copy and paste a German text from a text processor, a PDF document, or the e-mail agent. Once the text is pasted into the application, the user is able to select any word in the text and do any of the following:

- look up word in the German English dictionary;
- request a Keyword in Context list of a selected word, retrieved from the German corpus;
- access information on the morphological structure; this function is implemented by transporting the user to the Canoo website. Websites are "surf-able". Users can follow links on the web page in order to access information they are specifically interested in;
- look up synonyms of a word in a thesaurus;
- look up common collocations of a word using the corpus;
- query the frequency with which the word occurs in the corpus;
- request the inflectional paradigm of the word; this is again implemented by creating a suitable query to the Canoo website.
- look up the word in the German Wikipedia

At any point, the user is free to select another word, either from the text that was imported or from the data retrieved from the database. Since the Wikipedia and Canoo pages are displayed in a real web browser, the user is free to use them as such and follow any hyperlink in the usual fashion.

The screen-shots in this chapter are intended to give the reader an idea of the functionality offered by QuickAssist and the information that is displayed to the user. The following paragraphs will provide an overview over the individual functions and the rationale behind including them in QuickAssist.

The *Englisch* button triggers a dictionary look-up. QuickAssist will first look up the word in the database containing wordform lemma pairs. If the form is found here, the

Figure 5.3: Wikipedia Function

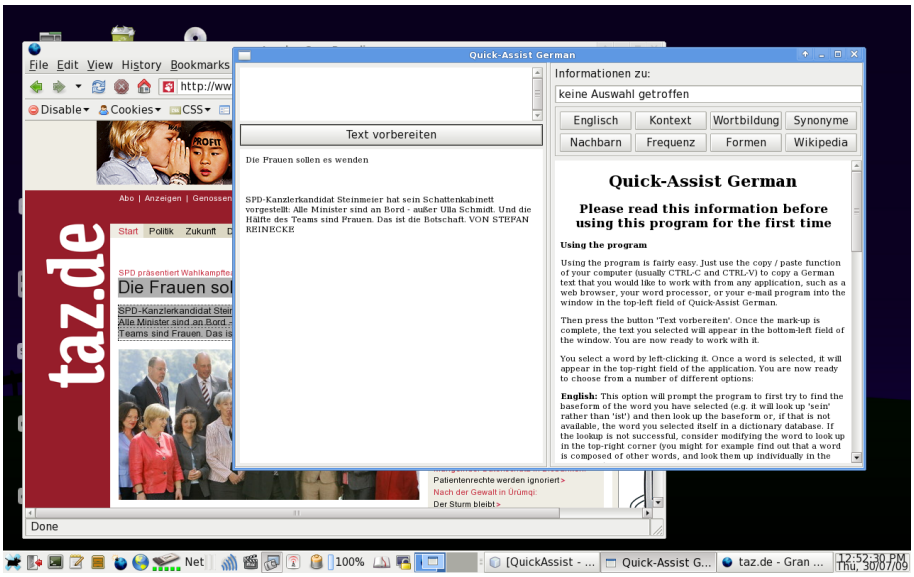


Figure 5.4: QuickAssist: Importing Text

matching lemma will be looked up in the German-English dictionary. As a fall-through option QuickAssist will look up the word form as it appears in the text.

This function was included to enable students to quickly look up a word that they

don't know. It was hypothesised that by enabling users to quickly determine the meaning of a word they will be able to process a German text quicker than by using traditional references. This in return should lead to a more positive reading experience and less frustration, because the reading process is not interrupted for too long and passages do not necessarily have to be reread, as it often happens using a printed dictionary.

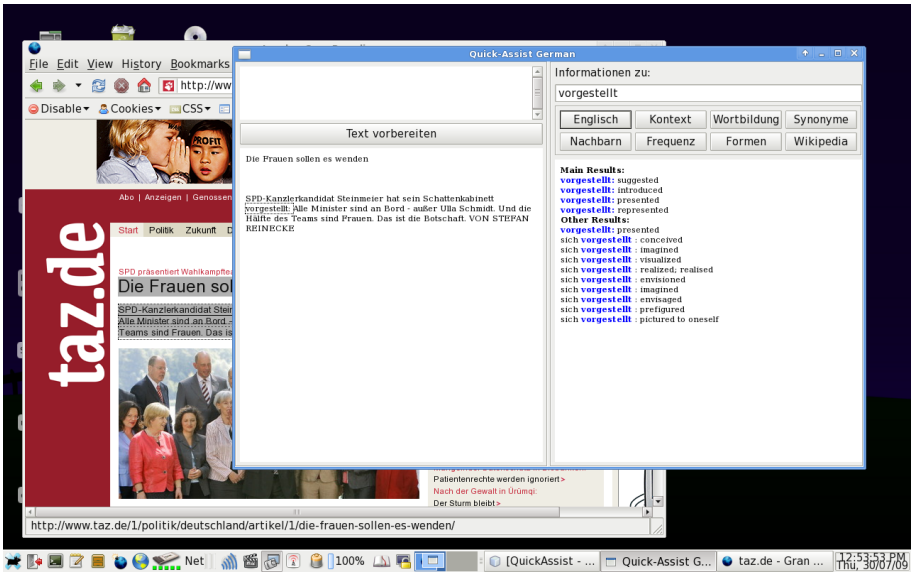


Figure 5.5: QuickAssist: German-English Translation

Nachbarn displays the most common neighbours of a wordform in the Wortschatz corpus. It was included to enable students to learn about common collocations of a word. These can act as indicators of registers in which a word is commonly used. It can also inform students about the semantic fields to which the word belongs. Most importantly, it can help students find out that the word might be a member of a multi word construction. QuickAssist is currently only able to look up single words (strings delimited by white space or punctuation marks). Therefore users will have to use a search engine to look up the multi-word construction or refer to a dictionary that is more exhaustive than the QuickAssist dictionary.

If a user presses *Kontext* QuickAssist generates a query to the Wortschatz database and

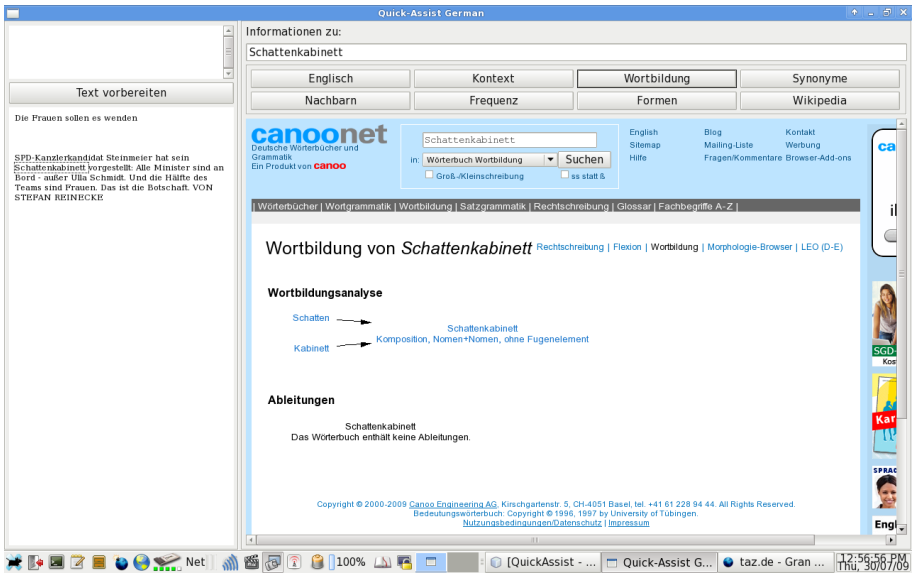


Figure 5.6: QuickAssist: Morphological Analysis using Canoo.net

retrieves a set of sentences that contain the same word form than the user has selected. These sentences are then displayed in the order they are retrieved from the database with the keyword highlighted. This function is intended to show users the word in a wide range of different contexts. This will provide them with an understanding of when a certain word is appropriate to be used, what it might mean in different contexts, whether it is a member of an idiomatic expression etc. The amount of information that users can gather on the basis of such keyword-in-context look-ups is of course largely dependent on the quality of the corpus and the range of texts it covers. QuickAssist comprises mainly newspaper articles, which clearly limits the range of possible text sorts, on the other hand, it is general enough to provide users with an idea of how a word is used in standard written German.

Frequenz will display the number of occurrences of the selected word form in the Wortschatz corpus. This was included to enable users to research whether a certain word they are planning to use in their own texts is an adequate choice. A dictionary will often

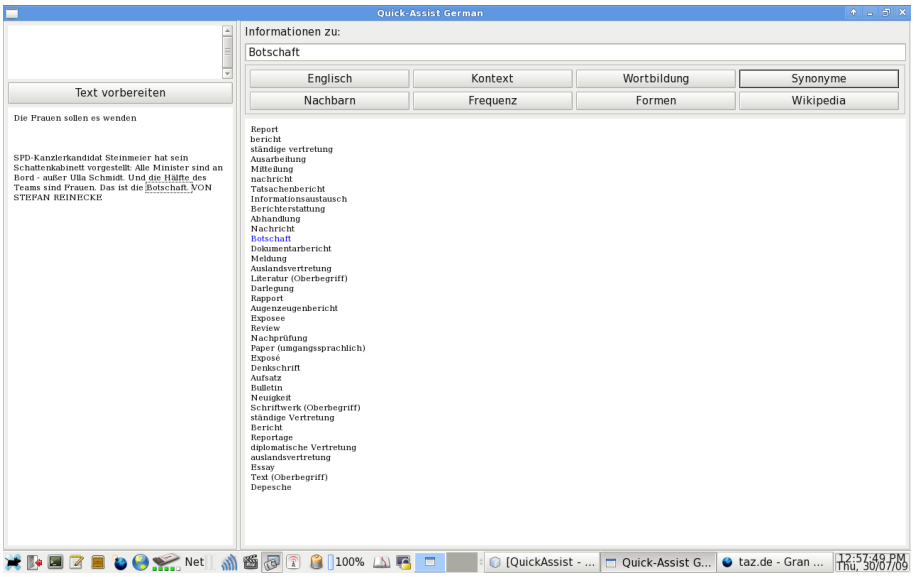


Figure 5.7: QuickAssist: Synonyms Function

provide a number of different translations for a given word without indicating directly which one of them will be the most adequate translation. Based on the fact that QuickAssist uses a corpus of contemporary German, users should generally find that a word that might be considered archaic by most speakers of German will have a lower frequency in the corpus than a translation that is more contemporary. Thus, if a user who is looking for an adequate translation of head (the body part) finds two possible translations: *Kopf* and *Haupt* in her dictionary, she can use the *Frequenz* function and will find that *Kopf* has a far higher frequency than *Haupt* and is likely the better choice.

As pointed out in chapter 2.4, it is necessary to understand the morphological structure of German words in order to determine their syntactic status and semantics. Especially compounds can be troublesome for learners of German, since no white space is used to delimit the individual constituents. If a learner does not know any part of a complex word, he will have a hard time determining which words to look up in a dictionary if the complex form is not listed by itself. *Wortbildung* generates a redirect request in the right

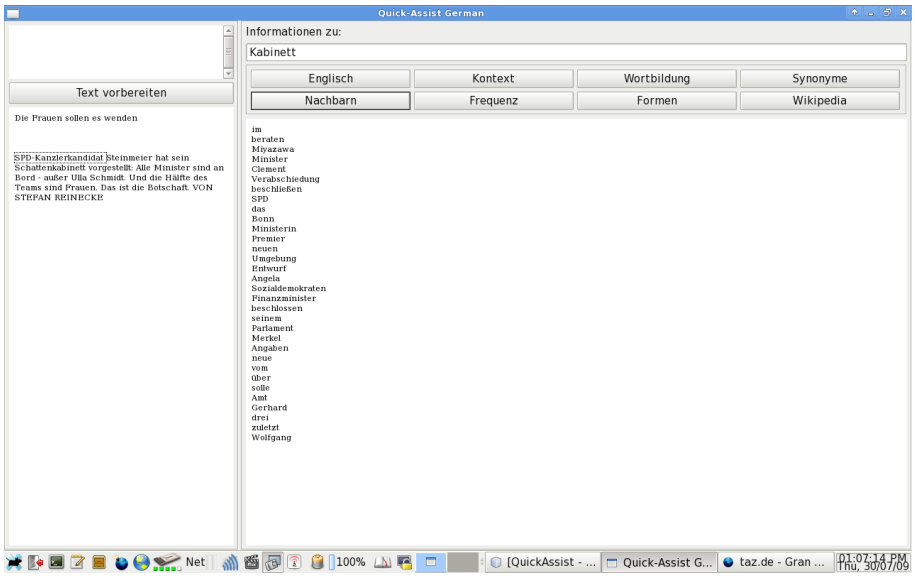


Figure 5.8: QuickAssist: Display of Direct Neighbours

browser module which calls the Canoo.net web page together with the request to analyse the selected word form. The next chapter contains an illustrated example of how users can use the information provided to look up the individual constituents of a complex word and determine its meaning.

Formen calls the Canoo.net website again. This time it tries to retrieve information on the inflectional paradigm that the selected word form belongs to. It was included to enable users to look up forms that might not be correctly lemmatised by QuickAssist, given that it has a limited number of wordform–lemma pairs stored in its database.

Checking *Synonyme*, users can look up synonyms of the selected word. Based on this information alone, together with the context of the word they are researching, they might be able to determine its meaning. The function might also prove helpful in case an English translation is not found in the dictionary. Provided the synonym function returns a suitable synonym, this can be looked up in the dictionary which might provide some

helpful information.

While the *Wikipedia* button was extremely popular in the user study, where participants at least originally tried it on all sorts of words, it was included for one particular purpose. While the dictionary and the corpus will usually provide users with a wealth of information on "normal" words, many texts that users are interested in will contain the names of people, places, institutions, events, etc. that will not be found in the dictionary or corpus, because of the fact that both of these resources were assembled before these terms became relevant. For example, looking up the name "Westerwelle" in the corpus will not return any useful information. Guido Westerwelle who is currently the Minister of Foreign Affairs in Germany entered the political stage some time after the corpus that QuickAssist uses was assembled. The German Wikipedia is a rich source for up-to-date cultural information with all of the quirks and inadequacies of its English counter part, but it will provide users with enough information to make sense of a certain concept that they are not familiar with, and it gives them more opportunities to read, explore and research. In the user study described in the next chapter, the picture of a famous German poet on the pertinent Wikipedia page was enough for users to determine that the text they had to read deals with a person who is already deceased, which they needed to know in order to understand the text adequately.

Chapter 6

User Study

A qualitative user study was considered the best way to get an understanding of the affordances of the application for the intended user groups: learners and instructors of German. While a quantitative study could be used to measure the effect of using the application on the receptive and productive vocabulary of users, only a qualitative study is able to show how users unfamiliar with the application approach it, and learn to work with it to process a text in order to understand it. A survey form handed out to a substantial number of users would only be able to capture their subjective assessment of this process. What I wanted to find out was, whether users were able to use the program effectively and whether it can enable them to read a text for understanding that was judged to be too difficult as to be understood without any additional help. Moreover, a detailed study of how users work with the program and what problems they encountered was also intended to determine shortcomings and technical bugs.

The purpose of this experiment was to study the users' learning experience using QuickAssist. This was the first time that the software was actually used by learners of German. Hence, some questions which participants were asked had the primary function to show what improvements are necessary to make the software more user friendly. It

was hypothesised that providing students with QuickAssist and some training on how to use it effectively would have an effect on their L2 reading experience. Modern language classrooms tend to devote little time to form-focused instruction and teaching new vocabulary items, concentrating instead on communicative skills. Students are often required to deal with the learning of German vocabulary and morphology on their own. Providing them with a tool to facilitate this self-directed learning process was hoped to enable them to improve in these areas.

Users of QuickAssist are able to import a German text of their choice and look up word forms in a bi-lingual dictionary, or a sizeable German corpus to see the word form in a variety of different contexts. The program can also provide statistical information a word's frequency, and on its neighbours to help users decide whether a certain word is good choice in a specific context. With the help of questionnaires and user walkthroughs, I intended to establish how students benefit from using the software and how these benefits can potentially be increased.

The study outlined below received Ethics Clearance (ORE# 14877). In keeping with the regulations for studies involving humans at the University of Waterloo, the questionnaires, screen capturing videos and audio recordings can only be used by the primary investigator. I am required to ensure that any reference to participants made during presentations or publications concerned with the findings of this study will be made in such a way that the identity of the participants remains undisclosed. Names of any participants will not be disclosed here, neither will many other details that could help in the identification of the study participants.

6.1 Student study

6.1.1 Student walkthrough

User walkthroughs (Hémard, 1999) were done with a group of four learners of German.

In a walkthrough, participants work with a software application. Depending on the setup of the study, they are either free to experiment with the software, or they are asked to complete tasks set by the researcher. During the walkthrough, participants are asked to report orally on what they are doing and to explain the rationale for their actions. In research literature, this methodology is often referred to as think-aloud-protocols. They have been used in psychological experiments in order to study cognitive processes. Because actions of participants can either be triggered consciously, but also without the participants being consciously aware of their actions, the reliability of this method has repeatedly been questioned. For a recent summary of this discussion, cf.: Hama & Leow (2010).

In the study outlined in this chapter, a mixed approach was used. Users where given time to experiment freely with the program to get an idea about its' functionality. Afterwards they were asked to complete a set task. The study was not designed to find out about subconscious or unconscious actions of the participants. The think-aloud protocols were used as a way for participants to document the strategies they were developing and using to accomplish certain tasks. As the walkthrough was their first exposure to QuickAssist, it can be safely assumed that using the functions provided by the program would happen with the participants being fully aware of it. There was simply no chance that automation could have taken place prior to the experiment. The same can of course not be said about the reading strategies participants were using.

I was given the opportunity to briefly introduce QuickAssist and the study during a regular class of a third year German course (see Appendix: 7.6). Students interested in the

study were provided with a letter outlining the study and a consent form (cf. Appendix: 7.6). Four students from this German class were willing to participate in the study. After a user walkthrough, described in more detail below, they were provided with my software free of charge. Four weeks later, I interviewed the students individually to find out how they had been using the software specifically and which problems they had encountered, in order to establish how students perceive the effectiveness of the software and what changes they thought needed to be implemented to improve the tool.

The exact structure of the user walkthroughs was as follows:

Phase 1 (5 minutes): Researcher demonstrates how the program works, what functions are available and answers possible questions.

Phase 2 (15 minutes) Warm-up: Student experiments with the program, explaining what he/she does and has the opportunity to ask the researcher for help.

Phase 3 (20 minutes) Task: Student is given a German newspaper article and questions on the contents of the text. While commenting on his/her actions the student tries to answer as many questions as possible with the help of the program.

Phase 4 (5 minutes): student comments briefly on his/her experience and reports on how he/she is planning to use the software in the next few weeks, giving reasons for this decision.

The learners were asked to experiment with the program for some time and were given the opportunity to ask questions about its functionality and on how to operate it. Then they were given the text (*Schiller GEZ article*, last accessed: 17 September 2010) shown below and the task to read it and summarise it orally in English. They were given twenty minutes to read the text with the help of the application. The entire session was audio

Aufgaben
1.) Lesen Sie den Text
2.) Fassen Sie den Text kurz zusammen
Schon GEZahlt, Herr Poet? Über 200 Jahre nach seinem Tod sollte der Nationaldichter Friedrich Schiller Rundfunkgebühren zahlen. Die GEZ verschickte Mahnbriefe an die sächsische "Friedrich Schiller"-Grundschule - und bemüht sich jetzt um Erklärung für die Panne.
Dresden - Die Briefe waren adressiert an "Herrn Friedrich Schiller" und ihre Aufforderung war deutlich: Schiller möge doch bitte Angaben zu seinem aktuellen TV- und Radiokonsum machen, hieß es darin.
In der "Friedrich Schiller"-Grundschule in Weigsdorf-Köblitz in Sachsen hielt man dies zunächst für einen schlechten Scherz. Doch einem Hinweis an die Gebühreneinzugszentrale (GEZ), dass der 1805 verstorbene Nationaldichter wohl nicht mehr in der Lage sei, ein Radio anzumelden, soll ein weiteres Mahnschreiben der Gebührenfahnder gefolgt sein.
Die GEZ bestätigte am Dienstag gegenüber der "Dresdner Morgenpost" die peinliche Panne - und bemühte sich um eine Erklärung. Man arbeite mit einer Riesenmenge an Daten, weshalb Fehler nicht sofort aufgedeckt würden. Zudem sei Friedrich Schiller kein so ungewöhnlicher Name, sagte eine Sprecherin.
Die Anschrift von "Friedrich Schiller" erhielt die GEZ nach eigenen Angaben von einem Adressenanbieter für die Zielgruppe "Haushalte".

Figure 6.1: Text used in the user study

recorded. In addition, a screen capture software was used to record the computer screen. This made it possible to study users' actions and their comments in detail.

The phase to experiment freely with the application was intended to give the participants an idea of the functionality that each of the buttons provided. In addition the text displayed by the application on start up is designed to inform users about the intended use of the functions and what output to expect.

What follows, is the summary of the four user walkthroughs. Some information is first given on the participant in order to give readers an idea of their language level, learner type, motivation, etc. In order to guarantee the anonymity of my study participants, I have chosen to not disclose their gender. I will be using the male forms of pronouns in the following description.

6.1.1.1 User One

User 1 is a retired Humanities professor. He has decided to take German courses for personal reasons. His main interest is German literature and poetry. He considers himself an avid reader, but as not very computer literate.

Five minutes of the warm-up phase are used to study the instructions and descriptions QuickAssist provides for the individual functions. Afterwards the user finds a text on a German city with the help of the German version of the Google search engine.

The user explores the functions of the program in some detail. He comments on the size of the font which is considered too small. The context-sensitive mouse-over help that is provided for all buttons is considered helpful.

The context function also impresses the user who perceives it as "a dictionary with loads of quotes." He also likes the information available on the Canoo website, but dislikes the fact that there is no back button that would make it possible to navigate back to

pages previously viewed.

After twenty minutes, user one starts working on the task. The user looks up relatively few words. The first word looked up is *anmelden*. Checking the context does not provide enough information to guess the word. The user decides to look it up in the dictionary which is successful. The next word is *Mahnbriefe*. Here, looking up the word in the dictionary fails. The user tries the word formation function and then looks up*mahnen*. Now the user is able to infer the meaning of the compound. Other words looked up are: *GEZ, Gebührenfahnder, Anschrift, erhielt, Adressanbieter, Zielgruppe, Haushalte*.

While selecting suitable functions does not seem to be a problem, the user clearly has problems with using the computer. Using the mouse and keyboard is slow. The user also comments that it is hard to find where he left off in the text when using one of the function buttons to look up information.

The user is able to summarize the content of the article in his own words at the end of the experiment.

The functionalities the user intends to use in the next four weeks are the dictionary and maybe the neighbours function, synonyms and frequency. The users thinks that the sentences provided with the context button are probably too difficult to be of much use. The user says that he will use the Wikipedia button "only if I have too much time"

In the interview carried out four weeks after the walkthrough, user one reports that he had great difficulty installing the program on the computer at home. He was also disappointed by the dictionary, who he feels he has outgrown. There seems to be a discrepancy between the intended use of the program (intense language study) and the coverage the dictionary offered. Although the user acknowledges that it is possible to analyse words into constituents and look them up individually, he says that he did not use this function. The results of such an analysis, according to the user, are imprecise because compounds can always have an idiomatic meaning. In addition, the user found that looking up words

on the Canoo website often did not work.

6.1.1.2 User Two

User Two is a third year German and French student in the early twenties. He is interested in the German community in Kitchener-Waterloo and works for one of the German clubs in the region. He also works three other jobs and helps a German superior who is German with his English correspondence. His siblings also learn languages at school and the study participant helps them with their homework.

The warm-up phase is relatively short. The user spends about three minutes reading the instructions and then about one and a half minutes to find a German text on the internet on the Oktoberfest. The first word looked up is "Geiersturzflug". He comments, "this is a long one, so I try Wortbildung". Word formation does help the user to find out the individual parts of the word, but even looking them up individually does not help the user to infer the meaning of the word as a whole. Only using the Wikipedia function finally reveals that the word is the name of a German band. Using the word formation function on the word *Schürzenjäger* the student comments, "this is really good for compound words." Using the forms function on another word, the user comments, "so I can use it if I write an essay." About the Wikipedia function the participant says, "this finds more cultural things."

The user starts working on the set task after twelve minutes. The first word looked up is *Nationaldichter*. The student tries functions randomly, often not leaving the program enough time to provide any output. The user finally uses the word formation function and decides to look up *dichter* in the dictionary. The resulting output is the translation of the adjective *dicht* which does not proof very helpful. It takes about three minutes until the user decides to look up the capitalized form *Dichter* which provides the information that leads the user to conclude that the word means national poet.

The user now begins to work more systematic. Encountering the word *GEZahlt*, he comments, “this is a little pun. I have to look this up on Wikipedia.” The user reads the entry on GEZ and asks for a way to get back to the previous screen. Other words looked up by user two include: *Einzug, Aufforderung, Scherz, Gebührenfahnder, Riesenmenge*. In most of the cases, using the word formation function helps the user to infer the meaning who now automatically capitalizes nouns before looking them up.

The user is able to provide the main idea of the text at the end of the task phase.

The student plans on using the program regularly. The most useful functions, he thinks, will be the word formation function and the direct link to Wikipedia. The user comments, “I will use it. This way, I will not have to have too many open windows.”

After four weeks the participant reports that he has used the program regularly for the homework assignments together with other NLP tools (beolingus and leo.org) to “cross-reference”. User two also reports that Canoo timed out on numerous occasions. He has found the synonyms function to be very helpful. Of the Wikipedia function, he reports, “it gets me surfing. I wouldn’t read as much with a book. Homework take more time now, just because I am more thorough.” The participant also expresses that he thinks his reading comprehension has improved by using the tool. He also used the frequency function to find out which words to use in his own writing. Suggestions for improvement include the option to keep notes (“I have to copy and paste everything into a text editor”) and to also offer a direct link to leo.org.

6.1.1.3 User Three

User three is a third year student of German and business in his early twenties. He lived in Germany for a number of years before his family moved to Canada. His hobby is computer programming

The warm-up phase takes about nineteen minutes. The user reads the introductions for about four minutes and then finds a text on a German contemporary writer on the internet. The participant seems more interested in technical details than in working with the application. He asks, e.g., "Why are the words not converted to lower-case before the look-up?" or, "is there an option to print?"

Working on the set task, the student appears to use functions not so much to help him understand the text, but to see how the program deals with unexpected or problematic entries. It is necessary on some occasions to remind him what the task is and that there is a time limit. The participant looks up the following words: *GEZ, Auffordertung, Gebühreneinzugszentrale, Gebührenfahnder, Mahnschreiben, Nationaldichter, Adressanbieter, Anschrift, Zielgruppe*.

After thirty-six minutes, the student is able to tell me what the gist of the article is.

The user plans to use the program to look up paradigms of nouns and verbs which he sometimes struggles with, but also for reading. He is not sure what the use of the frequency button is. The neighbours function might proof useful in certain contexts, he says.

During the interview four weeks after the walk through, the user says that he did not have much time to use the program. He only used it twice and found it more convenient to use leo.org to work on his homework assignments.

6.1.1.4 User Four

User four is a retired teacher who has come back to university to study German out of personal interest. He has a German background, visits Germany frequently and also researches his family's history.

The user studies the instructions for about two minute at the beginning of the warm-

up phase. Then he looks for a suitable text to practice with. He finds an article on computers. The first word the user looks up is *Einsatzbereiche*. He uses a number of different functions on it. Using word formation, he comments on Canoo.net, "I actually used this before." Looking up *Waffen* with the context function he comments, "This will be useful for me to see how the word is used in a sentence." On seeing the semantic network that the word formation function provides for the same word he says, "Oh, there is a lot of information here."

The participant starts working on the main task after fourteen minutes. The first word he looks up is "sei", because, he explains, his class is discussing the Konjunktiv I at the moment. He uses the paradigm function, looks at the output and confirms, "Ya, it's all there." The words the user looks up include: *GEZ, Gebühren, Aufforderung, Mahnschreiben, peinlich*. The user expresses a few times that he finds the text too hard to understand and that he would need more time to work with it. Nevertheless, he is able to tell me at the end of the walkthrough what *peinliche Panne* referred to.

The user expresses at the end of the experiment that he is keen to use the software on his own and that he wants to use it for his homework and for his research of his family's history.

After four weeks, the student confirms that he did use the program frequently for his readings. He says that the dictionary "is as good as I expected." He also says that he used the Wikipedia function frequently, because, "pictures are often better than words." He also says that he used the context function to find out how words are used "in modern days." He also likes the word formation function because, "I remember words better if I know what their components are." About the frequency function he says that he did not use it, because, "I am a learner, not a researcher."

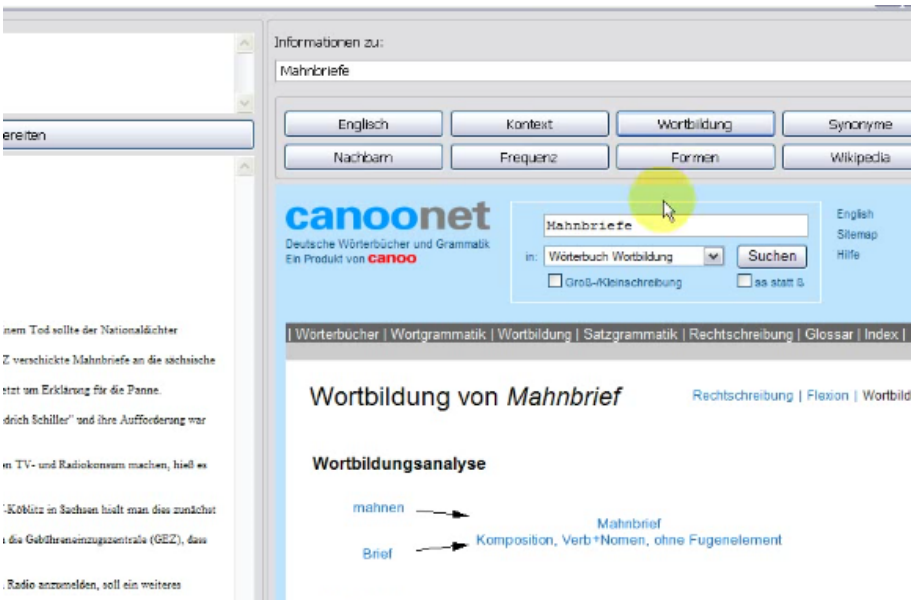


Figure 6.2: Morphological analysis of *Mahnbrief*

6.1.1.5 Findings

The text selected for the task was considered difficult by all students. The fact that many of the vocabulary items where looked up by all four students suggests that all of them were not familiar with a fair number of important key words. Given the number of words looked up by the subjects, it is also clear that the amount of unknown words clearly exceeded the two or three percent of unknown vocabulary items that learners according to current beliefs should be able to handle independently.

All of the students were able to developed strategies that enabled them to analyse unknown compounds and look up the constituents individually if the compound is not listed in the dictionary.

Figures 6.2 and 6.3 illustrate this. In the first figure, we can see that the participant has decided to use the morphological analysis feature of QuickAssist to learn about the structure of *Mahnbrief* which is not listed in QuickAssist's dictionary. Using this information

the participant decides to look up the infinitive form *mahnen* in the dictionary. Together with the information that *Brief* means letter in English, the participant is able to infer that *Mahnbrief* is a reminder for late payment.

All of the students were able to summarize the main idea of the text. While most of them struggled with the last sentence, which was considered hard, because it has unproportionally many unknown words and is arguable disconnected from the remaining text, all of them where able to tell me what the "mishap" was, what the GEZ is and that Germans apparently have to pay for watching TV, a fact that none of the participants was aware of at the beginning of the walkthrough.

I had the impression that the two mature students proceeded slower in their reading and tried to work through the text in a sequential manner, looking up every word that they did not know. The two younger participants seemed to use skimming as a technique. While they progressed faster they had to reread passages later when they tried to answer the questions on the text. On the other hand, all participants, used the program on familiar words during the warm-up phase to test the program, to see, I would hypothesize, whether the output of the program was what they expected.

To sum up then, one student did not find the time to use the application at all. Another student had an intense dislike for computers, became frustrated quickly with the program because of technical issues and because she tried to use it to read classical German poetry. As both the dictionary and the corpus were designed with modern German in mind, the resources QuickAssist offers proved inadequate for the task.

The two other students reported mainly positive experiences. One had used the application for research on his family's history. The other used the program alongside other resources such as online dictionaries. She was in fact the only participant that made use of the other features of the program because, as she reported, she did not only use it for receptive purposes, but also to produce German texts. Writing her own texts, she found

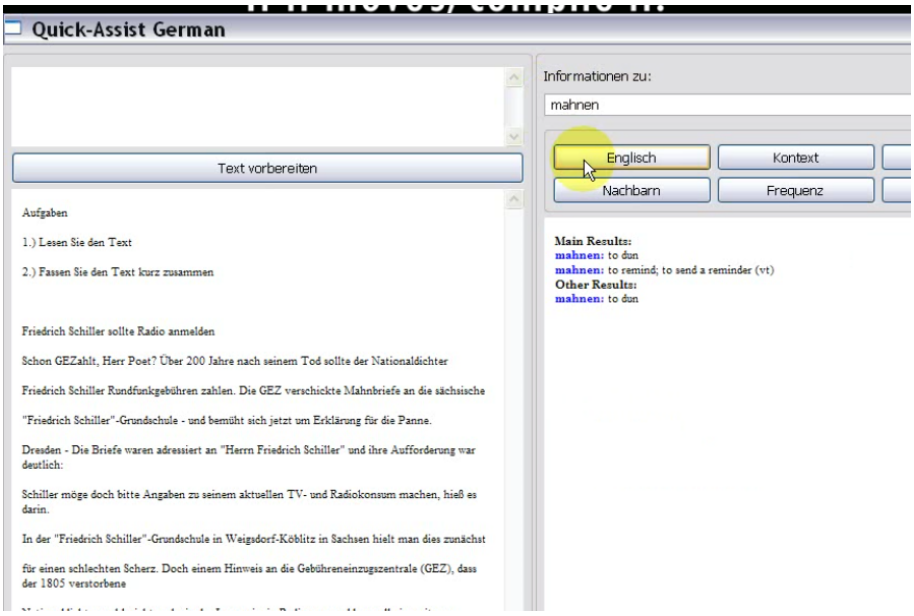


Figure 6.3: Look-up of *mahnen*

it useful to check word frequencies in the corpus, collocations and also information on inflectional paradigms of words.

Both of the students who reported that they had used QuickAssist in the weeks before the interview declared that they would continue to work with the application in the future since they found it helpful for their reading. They said that while the dictionary of Quick-Assist was fairly limited compared to other online resources, it enabled them to read a text quickly and with less effort than using traditional references or online dictionaries.

6.2 Instructor study

A group of three German instructors, consisting of two teaching assistants and a professor of German were also asked to evaluate the software. An adapted form of the ISTE (International Society for Technology in Education) software evaluation (cf. Appendix: 7.6)

form was used to assess the suitability of the software as a tool for learners to improve their German and as a tool for instructors to help create form-focused language exercises. While there are several different software evaluation forms available, the ISTE form is described as "a highly regarded general software evaluation form for educators," (Hubbard, 2006). It was selected for this study because it is designed to cover a wide range of educational software and questions are general enough for testers to express their opinions in a fairly unconstrained manner. A few extra questions were added to the ISTE form that are concerned with the suitability of QuickAssist for instructors and possible improvements to the software.

Instructors were recruited by distributing a letter to all German teaching faculty and teaching assistants (cf. Appendix: 7.6).

The structure of the sessions with individual instructors was planned as follows:

Phase 1 (10 minutes): Researcher demonstrates how the program works, what functions are available and answers possible questions.

Phase 2 (15 minutes): Instructor experiments with the program, explaining what he/she does and has the opportunity to ask the researcher for help.

Phase 3 (30 minutes): Instructor decides on a text that he/she could possibly use during one of his/her next classes. With the help of the instructor interface the instructor prepares auto-generated exercises, adapts them to his/her specific needs and comments on the working process.

Phase 4 (5 minutes): Instructor comments briefly on his/her experience with the application and reports on whether he/she would be using the software and for what purposes, giving reasons for this decision, filling in the attached questionnaire with the help of the researcher.

The instructors were given the program to experiment with. They were asked to comment on what they were doing with the program while a screen capturing tool recorded the desktop. Afterwards they were asked to rate the program using a modified version of a standard learning software evaluation form (7.6) and to comment on possible applications for learners and instructors.

The detailed results of the instructor study can be found in 7.6.

The comments of the instructors found that the application can speed up the understanding of a text, provided that the user had an adequate level of German and that the text was not too complicated. Mainly because of the nature of the corpus, it was thought in general that the user had to be an advanced learner of German to profit from using the program. In its current form, instructors could not imagine that they would be using the program themselves, to create exercises, for example, but that it might also be useful for other groups of users like translators. While the instructors were largely impressed with the features, they were pessimistic when asked whether they thought if the program could help with the acquisition of vocabulary.

6.3 Results

The study found that all four student participants were able to use the application with very little initial training to work successfully on a task that involved answering comprehension questions on a German text that contained a number of low frequency vocabulary items, complex compounds, and the names of persons and institutions that the students were not familiar with. Two of the participants also reported after four weeks that they had used the tool successfully for the completion of assignments in their German courses or even for individual research.

One of the more interesting results of this small scale study, I find, is that modern

learners are obviously able to quickly learn to work with a piece of software fast and efficiently. None of my participants reported that they found the tool too complicated to use. They discovered a number of ways in which the software could be improved and were able to clearly identify the capabilities as well as the limitations of the program. Some of the suggestions for improvements were:

- Adding the ability to look up multi word constructions.
- Adding a "back button". The subjects welcomed the option to browse for interesting information, using the program much like a web browser. In order to return to previously viewed pages, a back button would be necessary.
- Some of the study participants expressed that it would be good if users could customise the appearance of the application with respect to colours and font sizes.
- The dictionary was found to be limited. While it enables users to quickly look up a word, it does not offer a coverage that is comparable to other freely available online dictionaries, like leo.org or beolingus.org.
- The corpus was considered useful, but all participants agreed that the fact that it mainly comprises newspaper articles is a drawback.

It remains to be shown, but it can be hypothesised that the exposure to electronic media, the familiarity with the internet and the fact that the computer has become part of everyday life, has had the effect that most of us have developed strategies to filter information, assess the quality of sources, the suitability of resources for a specific task, and others. There also seems to be a difference in the use of strategies when it comes to age. In the user walkthroughs, there were two students in their early twenties and two retired persons who took the German course out of interest. While both mature learners were trying to read and comprehend the text in a linear fashion, the younger learners

used skimming techniques and were also somewhat more selective with the use of the program's functionalities and hence able to answer more of the comprehension questions.

Chapter 7

Conclusions

In this chapter, I would like to return to the research questions that are outlined in the introduction and revisit them one by one.

7.1 Question 1

> What can a software application look like that can potentially be used to help learners of a foreign language—and more specifically, learners of German as a foreign language—to extend their active and passive vocabulary, deepen their insight into the systematic rules that govern German word formation, and to improve their reading comprehension skills, and how does this sofware fit within current CALL applications, CALL theory and practice?

I argue in this dissertation and elsewhere (Wood, accepted for publication in 2011, submitted) that dedicated tutorial CALL software is currently only able to help language learners at the beginner level.

This is because of the limitations that available CALL technologies place on the learning situation. Only because tutorial CALL software controls learning materials and con-

strains admissible user input, it is able to check this input for correctness and provide adequate feedback.

Using these technologies, beginning learners are able to acquire a basic working vocabulary. The dissertation has shown that the concept of vocabulary knowledge is complex. Most available vocabulary drill software uses the concept of flash cards (cf. for example Wood, 2010). This necessitates that the 'meaning' of a word is reduced to one or possibly a few possible translations. The knowledge that speakers have of this word comprises far more. In addition to a range of semantic meanings that they have stored in their mind, they know about it's syntactic use, constraints with regard to registers of speaking and jargons, its internal structure, and other aspects.

In order to enable students to acquire a word in this complex sense, they will have to be exposed to it in multiple contexts and have the opportunity to study all aspects of the word. One of the most efficient ways to provide this degree of exposure to vocabulary, it is argued, is extensive reading.

The computer can function as a facilitator for students who have moved past the beginning and lower intermediate stages of their acquisition process and wish to study authentic texts of their choice in the foreign language. While it is possible to use traditional reference works for independent reading, this usually means that texts have to be at a difficulty level close enough to what students would be able to cope with without the help of reference material. If a text contains too many unknown words, or words used in a way unfamiliar to what the learner is used too, the time spent on looking up individual words will rapidly result in learners getting too frustrated and giving up.

With QuickAssist and similar software, students can look up words faster and this results in a more fluent reading experience. In addition, this software makes available a range of other tools that provide students with information that is not usually provided in a dictionary. In the case of QuickAssist, users can look up words in a corpus of contem-

porary German texts in order to study the word in different contexts and get a better idea of the range of its semantic features, its syntactic use and common situations in which it is used. The corpus also makes it possible that the program can provide users with information about the word's frequency and common neighbours. This way, users can research whether a certain word is adequate in a given situation and can thus be used to build active vocabulary. It also provides a thesaurus that can be used to look up synonyms which in turn enables users to broaden the range of vocabulary used in their own productions. The software is able to quickly direct users to morphological analyses of words, which has proven to be helpful for learners of German to work out the meaning of complex compounds and derivations, but is also intended to give users an insight into the system underlying German word formation, as is the option to look up the inflectional paradigm of words. In order to provide users with up-to-date information on people, events and cultural artefacts, users can also look up words in the German version of the Wikipedia, a quickly growing encyclopaedic database which is second only to the English version (Source: Tagesschau, 24 September 2010).

7.2 Question 2

> Can the computer serve as a tool to assist learners to achieve their goals by providing them with a range of features that are intended to help them work with a text of their choice in the target language?

The user study presented in the last chapter, I would argue, provides strong support for an affirmative answer to this question. With a limited amount of training, students were able to use the tool efficiently for their own reading projects. While there are other, similar software applications that provide students with the ability to look up words, etc., QuickAssist remains the only one to date that enables students to work with any text of

their choice. The dictionary, the corpus, and the thesaurus, etc. impose some constraints on what sorts of texts can be worked with adequately, but those study participants who tried to use it to read prose in modern German for understanding had an overall positive experience.

7.3 Question 3

> Is it possible to develop an application with these capabilities that can be used in a classroom context, but that learners can also use independently?

So far, no research has been done to show whether the tool can be used in a classroom context. A few colleagues have expressed the wish to use the application in their language classes, and the tool will be made available to them on completion of this dissertation. However, given that QuickAssist provides an easy to use interface to a corpus of German texts, there is no reason to believe that it cannot be used in the same ways that corpora and concordancers have been used in data driven language learning for the last twenty years. It will remain to be seen, in what other ways instructors will make use of the tool in their courses.

7.4 Reflections on the development

I think that finding an adequate development paradigm for QuickAssist was vital for the success of the project. Adhering to Colpaert's ADDIE had the advantage that the design of the application was driven by current SLA theories pertaining to the acquisition of vocabulary and to learning a language in general. It helped to find a set of tools that had the potential to be useful to the particular type of language learners I had in mind at the

beginning of the project, learners of at an intermediate to advanced level, who want to work fairly independently and who are not catered for by mainstream CALL technology

From the technological perspective, I think, that I personally benefited from experimenting with a wide range of different programming languages, and technologies. More than three years after the first prototype was developed, QuickAssist still runs on at least two common platforms. If I had to do it all over again, I would probably not select Java and SWT, but this is largely because I am not sure what long-term effects the acquisition of Sun by Oracle will have on the development of open source applications. For the user study, I think developing QuickAssist with a distributed database, was the only feasible solution. In order to make the software widely available, however, a client server model would obviously be the better choice. I will return to this later.

I was able to show that the basic concept behind QuickAssist, to make NLP applications available to language learners via an intuitive interface, is a useful one. The technical implementation is manageable both in terms of work and financial resources. From an HCI perspective, of course, there is room for improvements.

QuickAssist, an ICALL application is able to deliver just-in-time and contingent information to learners that is relevant for their learning process in general and pertains directly to the learning situation. This is possible because of the use of available NLP applications and external resources. The re-use of linguistic applications that I have argued for here and in Wood (2008) is feasible. QuickAssist was developed largely using existing NLP applications. In return, some of its components have been included in other applications (Schulze et al., submitted). The software that I have developed can be released under the GPL as intended. The database necessary to use it, however, contains material that cannot be released under the same licence. Work is under way to correct this. See below.

I hope that this project together with the effort of many other open source developers

will contribute to the steady improvement of publicly available CALL software and to raise the awareness for the problems connected with closed source programs, copyrights and ever increasing infringements of the rights of users and developers alike.

7.5 Reflections on the study

The user study was able to confirm my initial hypothesis that learners profit from the use of NLP applications if they are able to use them through an interface that is intuitive for them to use. All learners were able to learn quickly how to analyse the meaning of complex words that they could not find in a dictionary. The follow up interviews also showed that QuickAssist can be used independently for extensive reading. Of course the fact that I did not work with pre- and post-tests does not make it possible to claim that there was a measurable increase in the active or passive vocabulary that can be attributed to the use of the application. However, since some participants have reported that they did use the application regularly over four weeks and were planning to continue using it gives me reason enough to claim that they will experience the benefits of extensive reading in a foreign language found in other studies. By providing them with tools that encourage them to to do form-focused work by themselves, and having been told that the subjects made use of these tools regularly, I also assume that this will lead to a noticeable long term effect.

As mentioned above, it remains to be seen in how far instructors will benefit from the tool. There are a number of colleagues who have expressed the wish to use QuickAssist once I release it, and I hope they will report about their experiences.

In order to find out whether the use of QuickAssist has a quantifiable effect on the learning process, a quantitative study with a control group will be necessary. I hope that my upcoming publications on the initial success of QuickAssist will be able to help me

secure a research grant to fund such a study.

7.6 Future plans

While QuickAssist is fully functional in its current form, it is not possible to distribute it freely together with the database in its current form. This is because the Wortschatz corpus is not available under the GPL. It is available to researchers for research purposes only. This has enabled me to use it for the development of the prototype used in the user study. Work is currently under way to find a suitable corpus that can be released together with the application under the GPL.

Finding another corpus will hopefully also address a few other issues that I am aware of. The Wortschatz corpus consists of a database of individual sentences. As it is not possible to determine the sentences that precede, or follow the current sentence, the maximal context that users can be presented with is a single sentence. Using a corpus comprising complete texts or excerpts would enable users to study words in larger context, which will provide them with more information and the opportunity for more reading. I also hope to be able to offer learners corpus samples that are suitable for their individual proficiency level. Currently there are two options that are being explored: one involves the use of different corpora (one each for beginners, intermediate learners, etc.), the other is the use of learner models that provide information on users' vocabulary knowledge based on their lookups. Finding adequate texts is possible with web searches. Researchers at the University of Tübingen are working on an intelligent search system that might prove useful for QuickAssist.

Two important features that are missing currently are planned to be implemented in the near future: looking up multiple word constructions and the identification of synthetic verb forms and the so-called separable verbs.

Most of QuickAssist's current limitations that I reported on in the last chapter can be addressed by changing the interface and the architecture:

Most of the participants in the user study found aspects of the application problematic that pertain to HCI, such as small fonts, wrong colour scheme, missing buttons, etc. I am currently planning on changing the interface to QuickAssist entirely and use a Firefox plugin like Apheios or a Adobe Flash interface. This would mean that the application would run in the user's web browser. Configuring QuickAssist would be achieved mainly by changing the browser's configuration. This would reduce the amount of code that needs to be written for QuickAssist to implement custom configuration of the interface.

This is of course only possible by changing the architecture of QuickAssist. I am planning on setting up a Apache Tomcat server in the near future that would act as server for the QuickAssist client. This has a number of advantages: The client would be far smaller in size and easier to distribute than QuickAssist in the current form. I can make it available as a simple download from the QuickAssist homepage. The dictionary, thesaurus, and the corpora can be updated or changed if a more suitable replacement is found. It will also be easier to use applications such as the Stanford Tagger and Tree Tagger, which have been excluded so far because they would have put too much demands on the users' systems in terms of memory or software environment. Users will immediately benefit from these changes. User models could be stored on the server and provided users give their consent and the project will receive ethics clearance, this data could be used in future studies.

And finally, I hope to be able to make QuickAssist available for other languages, as well. Currently I am applying for a grant to port QuickAssist to French. Given our enrolments, this would enable me to do a qualitative study as outlined above, a project that I could not hope to do with the small number of German students in our intermediate and advanced courses at the University of Saskatchewan.

I hope that QuickAssist will continue to improve and prove useful to many language learners and instructors and hope to find collaborators interested to realise these and other plans.

Appendices

The following documents were used for the recruitment of participants for the user study.

Appendix 1: Letter to Instructors

The following letter was distributed to German instructors in the department.

University of Waterloo

Date

Dear instructor*:* This letter is an invitation to consider participating in a study I am conducting as part of my PhD degree in the Department of Germanic and Slavic Studies at the University of Waterloo under the supervision of Professor Mathias Schulze . I would like to provide you with more information about this project and what your involvement would entail if you decide to take part.

The title of the study is: *Quick-Assist. The effects of using CALL software to facilitate the acquisition of German vocabulary and word formation rules on the learning outcome*.

As part of my dissertation project, I am developing a software that is designed to help learners of German to learn new vocabulary as well as German word formation rules. Users of my software will be able to import any (public-domain) German text of their choice and look up word forms in a bi-lingual dictionary and a sizable German corpus to see the word form in a variety of different contexts. The program can also provide information such as collocations, neighbours, and frequencies of word forms to help users decide whether a certain word is a good choice in a specific context. In addition, the program can generate exercises automatically to provide further learning opportunities. With the help of questionnaires and user walk-throughs, I will try to establish how students and instructors benefit from using the software and how these benefits can potentially be increased.

Participation in this study is voluntary. It will involve a user walk-through in which you are using the program with my help to create some exercises that could be used by German students in one of your classes. While you are doing this, I will ask you

to explain what you are doing and your reasons for your actions. Your voice will be recorded during this session. I will also use screen capturing software to record what you are doing on the computer. This walk-through will last about one hour. I will also ask you to fill in a questionnaire that is designed to asses the program The walk-through will take place at a mutually agreed upon location and time. You may decline to carry out any of the tasks in the user walk-through or to answer any of the questions in the questionnaire if you so wish. Further, you may decide to withdraw from this study at any time without any negative consequences by advising me. All information you provide is considered completely confidential. Your name will not appear in any thesis or report resulting from this study, however, with your permission anonymous quotations may be used. Data collected during this study will be retained for 3 years in a locked office in the Germanic and Slavic Studies department and then confidentially destroyed. Only I and my supervisor will have access. There are no known or anticipated risks to you as a participant in this study.

If you have any questions regarding this study, or would like additional information to assist you in reaching a decision about participation, please contact me at 519 584 1770 or by email at p2wood@uwaterloo.ca. You can also contact my supervisor, Professor Mathias Schulze at mschulze@uwaterloo.ca.

I would like to assure you that this study has been reviewed and received ethics clearance through the Office of Research Ethics. However, the final decision about participation is yours. If you have any comments or concerns resulting from your participation in this study, please contact Dr. Susan Sykes of this office at 519-888-4567 Ext. 36005.

I hope that the results of my study will be of benefit to German students and linguists researching the use of computers in the area of language learning.

I very much look forward to speaking with you and thank you in advance for your assistance in this project.

Yours sincerely,

Peter Wood

CONSENT FORM

I have read the information presented in the information letter about a study being conducted by Peter Wood of the Department of Germanic and Slavic Studies at the University of Waterloo. I have had the opportunity to ask any questions related to this study, to receive satisfactory answers to my questions, and any additional details I wanted.

I am aware that I have the option of allowing my walkthrough to be audio recorded to ensure an accurate recording of my responses.

I am also aware that excerpts from the interview may be included in the thesis and/or publications to come from this research, with the understanding that the quotations will be anonymous.

I was informed that I may withdraw my consent at any time without penalty by advising the researcher.

This project has been reviewed by, and received ethics clearance through, the Office of Research Ethics at the University of Waterloo. I was informed that if I have any comments or concerns resulting from my participation in this study, I may contact the Director, Office of Research Ethics at 519-888-4567 ext. 36005.

With full knowledge of all foregoing, I agree, of my own free will, to participate in this study.

YES NO

I agree to have my walk-through recorded with a screen capturing software and an audio recorder .

YES NO

I agree to the use of anonymous quotations in any thesis or publication that comes of this research.

YES NO

Participant Name: _____ (Please print)

Participant Signature: _____

Witness Name: _____ (Please print)

Witness Signature: _____

Date: _____

subsection*Appendix 2: Recruitment Script The following text was used to be read to students in order to find potential subjects for the student user walk throughs.

In Class Recruitment Script

Hello, my name is Peter Wood and I am a doctoral student in the Department of Germanic and Slavic Studies. I am currently working on my doctoral dissertation under the supervision of Professor Mathias Schulze. I have developed a computer program that is intended to help learners of German to improve their knowledge of German vocabulary and of German word formation rules. You will see a presentation of this software and you are all welcome to use it for the duration of this term. I would also like to invite you to take part in a study that is designed to find out whether this program helps students and instructors and what sort of improvements may have to be made. If you volunteer as a participant in this study, you will be asked to do a user walkthrough with me. We will find a suitable time at which we can meet in the PhD office in ML 242, I will show you how the program works, and will ask you to work with it for some time while you explain what you are doing with the program. Screen capturing software will record your actions and your voice will be recorded as well. This walkthrough will take about 1 hour. In about 6 weeks you will be asked to participate in an interview in which I will ask you how you have used the program in the meantime. This interview will only take about twenty minutes.

I would like to assure you that this study has been reviewed and received ethics clearance through the Office of Research Ethics. However, the final decision about participation is yours.

If you are interested in participating, please ask for an information letter and read the attached consent form. I am passing around a sheet on which I would ask you to write down your name and e-mail address. I will be contacting you to find a suitable time for a walkthrough with you. Before the beginning of the walkthrough we will go over

the consent form together and you will have the chance to ask any questions you have regarding the walkthrough before you sign the form. If you have any questions now or at a later time, please ask me now, after the class or just write me a quick e-mail.Thank you.

Appendix 3: Letter to Students

The following letter was distributed to German students interested to participate in the study.

University of Waterloo

Date

Dear student*:*

This letter is an invitation to consider participating in a study I am conducting as part of my PhD degree in the Department of Germanic and Slavic Studies at the University of Waterloo under the supervision of Professor Mathias Schulze . I would like to provide you with more information about this project and what your involvement would entail if you decide to take part.

The title of the study is: *Quick-Assist. The effects of using CALL software to facilitate the acquisition of German vocabulary and word formation rules on the learning outcome*.

As part of my dissertation project, I am developing software that is designed to help learners of German to learn new vocabulary as well as German word formation rules. Users of my software will be able to import any (public-domain) German text of their choice and look up word forms in a bi-lingual dictionary, or a sizable German corpus to see the word form in a variety of different contexts. The program can also provide information such as collocations, neighbors, and frequencies of word forms to help users decide whether a certain word is a good choice in a specific context. In addition the program can generate exercises automatically to provide further learning opportunities. With the help of a user walkthrough and a short interview, I will try to establish whether and how students benefit from using the software and how these benefits can potentially be increased.

Participation in this study is voluntary. It will involve a user walkthrough in which

you are using the program with my help to complete some tasks while you explain what you are doing. Your voice will be recorded during this session. I will also use screen capturing software to record what you are doing on the computer. This walkthrough will last about one hour. About six weeks afterwards you will be asked to participate in a short interview (about 20 minutes) in which I will ask you about the experiences you have had with the program. Both the walkthrough and the interview will take place in the German PhD office in ML 242 at a mutually agreed upon time. You may decline to carry out any of the tasks in the user walk through or to answer any of the interview questions if you so wish. Further, you may decide to withdraw from this study at any time without any negative consequences by advising me. All information you provide is considered completely confidential. Your name will not appear in any thesis or report resulting from this study, however, with your permission anonymous quotations may be used. Data collected during this study will be retained for 3 years in a locked office in the Germanic and Slavic Studies department and then confidentially destroyed. Only I and my supervisor will have access. There are no known or anticipated risks to you as a participant in this study.

If you have any questions regarding this study, or would like additional information to assist you in reaching a decision about participation, please contact me by email at p2wood@uwaterloo.ca. You can also contact my supervisor, Professor Mathias Schulze at mschulze@uwaterloo.ca

I would like to assure you that this study has been reviewed and received ethics clearance through the Office of Research Ethics. However, the final decision about participation is yours. If you have any comments or concerns resulting from your participation in this study, please contact Dr. Susan Sykes of this office at 519-888-4567 Ext. 36005.

I hope that the results of my study will be of benefit to German students and linguists researching the use of computers in the area of language learning.

I very much look forward to speaking with you and thank you in advance for your assistance in this project.

Yours Sincerely,

Peter Wood

[Warning: Draw object ignored]

CONSENT FORM

I have read the information presented in the information letter about a study being conducted by Peter Wood of the Department of Germanic and Slavic Studies at the University of Waterloo. I have had the opportunity to ask any questions related to this study, to receive satisfactory answers to my questions, and any additional details I wanted.

I am aware that I have the option of allowing my interview to be audio recorded to ensure an accurate recording of my responses.

I am also aware that excerpts from the interview may be included in the thesis and/or publications to come from this research, with the understanding that the quotations will be anonymous.

I was informed that I may withdraw my consent at any time without penalty by advising the researcher.

This project has been reviewed by, and received ethics clearance through, the Office of Research Ethics at the University of Waterloo. I was informed that if I have any comments or concerns resulting from my participation in this study, I may contact the Director, Office of Research Ethics at 519-888-4567 ext. 36005.

With full knowledge of all foregoing, I agree, of my own free will, to participate in this study.

YES NO

I agree to have my walkthrough recorded with a screen capturing software and an audio recorder .

YES NO

I agree to the use of anonymous quotations in any thesis or publication that comes of this research.

YES NO

Participant Name: _____ (Please print)

Participant Signature: _____

Witness Name: _____ (Please print)

Witness Signature: _____

Date: _____

Appendix 4: Instructor Questionnaire

The following questionnaire was used during the instructor walkthroughs. It is an adapted form of a standardized evaluation form for learning software.

Questionnaire for Instructors

(I) Instructional Design

Please rate (0-100%).

a) Use of this application promotes:

Creativity ___

Higher Order Thinking ___

Collaboration ___

Problem Solving ___

Discovery ___

Memorization ___

b) Please rate the validity of the following statements:

The student controls the pacing. ___

Use of the program stimulates curiosity. ___

Use of this program challenges the student. ___

The program offers real-world connections. ___

c) Strengths:

d) Weaknesses:

e) Describe the pedagogy incorporated in the design:

(II) Suitability

a) Is the application a suitable tool for **instructors**? Please explain your decision and name areas in which the program can be used / cannot be used providing reasons if possible.

b) Is the application a suitable tool for **learners** of German? Please explain your decision and name areas in which this program can be used / cannot be used, considering different proficiency levels of learners.

(III) User Interface

Please comment on the design of the user interface (fonts, sizes, labelling of buttons and menu items, availability and quality of online help, etc.)

(IV) Recommendations

(V) Your overall quality rating (0-100%) _____

Section I of this questionnaire is adapted from the Educational Software Evaluation Form available from the International Society for Technology in Education (ISTE). URL: http://www.iste.org

Appendix 5: Feedback Letter

The following letter was prepared for participants in the user study.

University of Waterloo

Date

Dear ***(Insert Name of Participant),***

I would like to thank you for your participation in this study. As a reminder, the title of the study was: *Quick-Assist. The effects of using CALL software to facilitate the acquisition of German vocabulary and word formation rules on the learning outcome*.

The data collected during interviews will contribute to a better understanding of how CALL applications can help in the language learning process.

Please remember that any data pertaining to you as an individual participant will be kept confidential. Once all the data are collected and analyzed for this project, I plan on sharing this information with the research community through seminars, conferences, presentations, and journal articles. If you are interested in receiving more information regarding the results of this study, you are welcome to read about them in my dissertation which will be publicly accessible in the Faculty of Arts in January 2009. If you have any questions or concerns, please contact me at either the phone number or email address listed at the bottom of the page.

As with all University of Waterloo projects involving human participants, this project was reviewed by, and received ethics clearance through, the Office of Research Ethics at the University of Waterloo. Should you have any comments or concerns resulting from your participation in this study, please contact Dr. Susan Sykes in the Office of Research Ethics at 519-888-4567, Ext., 36005.

Peter Wood

University of Waterloo

Germanic and Slavic Studies

519 584 1770

p2wood@uwaterloo.ca

Appendix 6: Instructor Study

Appendix 6.1: Answers Provided by Instructor One

Questionnaire for Instructors

(I) Instructional Design

Please rate (0-100%).

a) Use of this application promotes:

Creativity 90

Higher Order Thinking 70

Collaboration 90

Problem Solving 90

Discovery 90

Memorization 50-60

b) Please rate the validity of the following statements:

The student controls the pacing. yes

Use of the program stimulates curiosity. yes

Use of this program challenges the student. yes

The program offers real-world connections. yes

c) Strengths: users can work with any text, look up any word, texts can be authentic

d) Weaknesses: synonym function, too little explanations, takes some time to learn how to use the program effectively

e) Describe the pedagogy incorporated in the design: the program can be used as a part of a communicative language course

(II) Suitability

a) Is the application a suitable tool for **instructors**? Please explain your decision and name areas in which the program can be used / cannot be used providing reasons if possible.

instructors can find out background information on words, they can use it for teaching and research and to analyse student texts (error analysis)

b) Is the application a suitable tool for **learners** of German? Please explain your decision and name areas in which this program can be used / cannot be used, considering different proficiency levels of learners.

students with intermediate to advanced language skills can use the program independently, as can grad students. It can teach them how to work with texts. 101 and 102 students need guidance and should only use the program selectively.

(III) User Interface

Please comment on the design of the user interface (fonts, sizes, labelling of buttons and menu items, availability and quality of online help, etc.)

It's a matter of taste

(IV) Recommendations

improve thesaurus

(V) Your overall quality rating (0-100%) _____

80

Appendix 6.2: Answers Provided by Instructor Two

Questionnaire for Instructors

(I) Instructional Design

Please rate (0-100%).

a) Use of this application promotes:

Creativity 70

Higher Order Thinking 90 (advanced learners) / 65-70 (beginners)

Collaboration 80

Problem Solving 90

Discovery 80

Memorization 70

b) Please rate the validity of the following statements:

The student controls the pacing. yes

Use of the program stimulates curiosity. yes (in the beginning)

Use of this program challenges the student. yes

The program offers real-world connections. yes (except "Umgangssprache")

c) Strengths: any text, develops learning strategies, independent learning

d) Weaknesses: could make learners concentrate too much on individual words, beginners need guidance

e) Describe the pedagogy incorporated in the design: not communicative, maybe whole language

(II) Suitability

a) Is the application a suitable tool for **instructors**? Please explain your decision and name areas in which the program can be used / cannot be used providing reasons if possible.

can help non-native instructors with preparation of classes

b) Is the application a suitable tool for **learners** of German? Please explain your decision and name areas in which this program can be used / cannot be used, considering different proficiency levels of learners.

Too advanced for beginners. Use from year two and up

(III) User Interface

Please comment on the design of the user interface (fonts, sizes, labelling of buttons and menu items, availability and quality of online help, etc.)

font too small, should be more colourful

(IV) Recommendations

back-button, make more flexible in recognizing forms

(V) Your overall quality rating (0-100%) 80

Appendix 6.3: Answers Provided by Instructor Three

Questionnaire for Instructors

(I) Instructional Design

Please rate (0-100%).

a) Use of this application promotes:

Creativity 60

Higher Order Thinking 75

Collaboration 50

Problem Solving 80

Discovery 85

Memorization 25

b) Please rate the validity of the following statements:

The student controls the pacing. yes

Use of the program stimulates curiosity. yes

Use of this program challenges the student. yes

The program offers real-world connections. possibly

c) Strengths: simplicity, accessibility of different functions, functions are straight forward, context help (mouse over)

d) Weaknesses: only simple words not multi-word units can be processed, the tag button is useless, order of buttons seems arbitrary, button to get back to instructions is missing

e) Describe the pedagogy incorporated in the design: learner centred

(II) Suitability

a) Is the application a suitable tool for **instructors**? Please explain your decision and name areas in which the program can be used / cannot be used providing reasons if possible.

not sure, possible useful for non-native TA's

b) Is the application a suitable tool for **learners** of German? Please explain your decision and name areas in which this program can be used / cannot be used, considering different proficiency levels of learners.

yes. Beginning learners can profit from the dictionary and can use the program with texts from their text books

(III) User Interface

Please comment on the design of the user interface (fonts, sizes, labelling of buttons and menu items, availability and quality of online help, etc.)

simple, clean, not distracting

(IV) Recommendations

integrate tag button, add back button, provide contextual help for all buttons

(V) Your overall quality rating (0-100%) 80

References

Abeillé, A. (2003). *Treebanks : Building and Using Parsed Corpora*. Dordrecht ; Boston: Kluwer.

Aitchinson, J. (1994). *Words in the Mind. An Introduction to the Mental Lexicon* (2nd ed.). Oxford: Blackwell.

Albrecht, U., Dane, D., Fandrych, C., Grüßhaber, G., Henningsen, U., Kilmann, A., et al. (2008). *Passwort Deutsch. Kurs und Übungsbuch* (Vols. 1–5). Stuttgart: Klett.

Alpheios. (last accessed: 17 September 2010, September). Alpheios. Available from `http://alpheios.net`

Amaral, L. A. (2007). *Designing Intelligent Language Tutoring Systems for Integration into Foreign Language Instruction*. Unpublished doctoral dissertation, Ohio State University, Columbus.

Anglin, J. M., Miller, G. A., & Wakefield, P. C. (1993). *Vocabulary Development. A Morphological Analysis* (Vol. 58) (No. 10). New York: Monographs of the Society for Research in Child Development.

Arnold, K., Gosling, J., & Holmes, D. (1997). *The Java Programming Language*. Reading, Massachusetts: Addison-Wesley.

Baayen, H., & Lieber, R. (1991). Productivity and English Derivation: A Corpus-Based Study. *Linguistics*, *29*(5), 801–843.

Bach, M. (1986). *The Design of the UNIX Operating System*. Englewood Cliffs, NJ: Prentice-Hall.

Bauer, L. (1988). *Intoducing Linguisic Morphology*. Edinburgh: Edinburgh University Press.

Bauer, L. (1998). When is a Sequence of Noun + Noun a Compound in English? *English Language and Linguistics*, *2*, 65-86.

Bax, S. (2003). CALL – Past, Present, Future. *System*, *31*, 13–28.

Bergenholtz, H., & Mugdan, J. (1979). *Einführung in die Morphologie*. Stuttgart: Kohlhammer.

Bernstein, B. (1971). *Class, Codes and Control* (Vol. 1). London: Routledge & Kegan Paul.

Biemann, C., Bordag, S., Heyer, G., Quasthoff, U., & Wolff, C. (2004). Language-Independent Methods for Compiling Monolingual Lexical Data. In A. Gelbkuh (Ed.), *CICLing 2004, LNCS 2945* (pp. 217–228). Berlin, Heidelberg: Springer Verlag.

Blanchette, J., & Summerfield, M. (2008). *C++ GUI Programming with Qt 4*. Upper Saddle River, NJ: Prentice Hall Press.

Booij, G. (2005). *The Grammar of Words*. Oxford: OUP.

Borin, L. (2002). *What Have You Done for Me Lately? The Fickle Alignment of NLP and CALL.* (Paper given at the "NLP in CALL" pre-conference workshop of the EuroCALL conference 2002)

Boulton, A. (2010). Data Driven Learning: Taking the Computer out of the Equation. *Language Learning*, *60.3*, 534–572.

Bray, T., Paoli, J., Sperberg-McQueen, C., Maler, E., & Yergeau, F. (2000). Extensible Markup Language (XML) 1.0. *W3C Recommendation*, *6*.

Bresnan, J. (2001). *Lexical-Functional Syntax*. Oxford: Blackwell.

canoo net. (last accessed: 13 September 2010). canoo net. Available from `http://www.canoo.net`

Carstairs-McCarthy, A. (2002). *An Introduction to English Morphology: Words and their Structure*. Edinburgh: Edinburgh University Press.

Chapelle, C. (1997). CALL in the Year 2000: Still in Search of Research Paradigms. *Language Learning & Technology*, *1*(1), 19–43.

Chapelle, C. (1998). Multimedia CALL: Lessons to be Learned from Research on Instructed SLA. *Language Learning & Technology*, *2*(1), 21–39.

Chinnery, G. (2006). Emerging Technologies. Going to the MALL: Mobile Assisted Language Learning. *Language Learning & Technology*, *10*(1), 9–16.

Chomsky, N. (1957). *Syntactic Structures*. The Hague: Mouton.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge,MA: MIT press.

Chomsky, N. (1995). *The Minimalist Program*. Boston: MIT press.

Chun, D., & Plass, J. (1996). Effects of Multimedia Annotations on Vocabulary Acquisition. *Modern Language Journal*, *80*(2), 183–198.

Colpaert, J. (2004). *Design of Online Interactive Language Courseware: Conceptualization, Specification and Prototyping. Research into the Impact of Linguistic-Didactic*

Functionality on Software Architecture. Unpublished doctoral dissertation, University of Antwerp, Antwerp.

Cook, V. (2003). Linguistics and Second Language Acquisition: One Person with Two Languages. In M. Aronoff & J. Rees-Miller (Eds.), *The Handbook of Linguistics* (pp. 488–511). Oxford: Blackwell.

Cook, V., & Newson, M. (2007). *Chomsky's Universal Grammar: an Introduction* (3rd ed.). Oxford: Blackwell.

Coxhead, A. (2000). A New Academic Word List. *TESOL Quarterly*, *34.2*, 213-238.

Creutz, M., & Lagus, K. (2002). Unsupervised Discovery of Morphemes. *Proceedings of the ACL-02 Workshop on Morphological and Phonological learning*, *6*, 21–30.

Culicover, P. (2009). *Natural Language Syntax*. Oxford: OUP.

Culicover, P., & Jackendoff, R. (2005). *Simpler Syntax*. Oxford: OUP.

Curbera, F., Duftler, M., Khalaf, R., Nagy, W., Mukhi, N., & Weerawarana, S. (2002). Unraveling the Web Services Web: An Introduction to SOAP, WSDL, and UDDI. *IEEE Internet Computing*, *6*(2), 86–93.

DeKeyser, R. M. (2001). Automaticity and Automaticaton. In P. Robinson (Ed.), *Cognition and Second Language Instruction.* Cambridge: CUP.

Dokter, D., Nerbonne, J., Schurcks-Grozeva, L., & Smit, P. (1998). Glosser-RuG; A User Study. In S. Jager, J. A. Nerbonne, & A. van Essen (Eds.), *Language Teaching and Language Technology* (pp. 149–166). Lisse,NL: Swets&Zeitlinger.

Donalies, E. (2007). *Basiswissen Deutsche Wortbildung*. Tübingen: A. Francke.

Duden. Die deutsche Rechtschreibung (21st ed., Vol. 1). (1996). Mannheim: Duden Verlag.

Duden. Die deutsche Rechtschreibung (25th ed., Vol. 1). (2009). Mannheim: Duden Verlag.

Dörnyei, Z., & Schmidt, R. (Eds.). (2001). *Motivation and Second Language Acquisition*. Honolulu: Second Language Teaching & Curriculum Center, University of Hawai'i.

Dörnyei, Z., & Skehan, P. (2003). Individual Differences in Second Language Learning. In C. J. Doughty & M. H. Long (Eds.), *Second Language Acquisition.* Oxford: Blackwell.

Eisenberg, P. (1985). *Grundriss der deutschen Grammatik*. Stuttgart: Metzler.

Eisenberg, P. (1998). *Grundriss der deutschen Grammatik: Das Wort*. Stuttgart: Metzler.

Ellis, N. C. (2002). Frequency Effecs in Language Processing. A Review with Implicaitions for Theories of Implicit and Explicit Language Acquisition. *SSLA*, *24*, 143–188.

Ellis, N. C., & Larsen-Freeman, D. (2006). Language Emergence: Imlications for Applied Linguisics. *Applied Linguistics*, *27/4*, 558–589.

Ellis, R. (2008). *The Study of Second Language Acquisition*. Oxford: OUP.

Felix, U. (2005). Analysing Recent CALL Effectiveness Research. Towards a Common Agenda. *Computer Assisted Language Learning*, *18*(1 + 2), 1–32.

Fleischer, W., & Barz, I. (2007). *Wortbildung der deutschen Gegenwartssprache*. Tübingen: Niemeyer.

Flynn, J. R. (2008). *Where have all the Liberals Gone?: Race, Class, and Ideals in America*. Cambridge: CUP.

Freedict. (last accessed: 13 September 2010). Sourceforge. Available from http://prdownloads.sourceforge.net/stardict/stardict-freedict-deu-eng-2.4.2.tar.bz2?download

Gardner, R., & Lambert, W. (1972). *Attitudes and Motivation in Second Language Learning*. Rowley, Mass.: Newbury House.

Garside, R. G., Leech, G. N., & McEnery, T. (1997). *Corpus Annotation: Linguistic Information from Computer Text Corpora*. London, New York: Longman.

Gass, S., & Selinker, L. (2008). *Second Language Acquisition: An Introductory Course* (3rd ed.). New Jersey: Lawrence Erlbaum.

Gazdar, G., & Mellish, C. (1989). *Natural Language Processing in PROLOG: An Introduction to Computational Linguistics*. Reading, Massachusetts: Addison-Wesley.

General Public Licence. (last accessed: 13 September 2010). Available from http://www.gnu.org/licenses/gpl.html

German Wikipedia. (last accessed: 13 September 2010). Wikipedia. Available from http://www.wikipedia.de/

German Word Frequency List. (last accessed: 17 September 2010). Wortschatz - University of Leipzig. Available from http://wortschatz.uni-leipzig.de/html/wliste.html (Retrieved on: July 12 2010)

Giegerich, H. (1999). *Lexical Strata in English: Morphological Causes, Phonological Effects*. Cambridge: CUP.

Glück, H. (Ed.). (1993). *Metzler Lexikon Sprache*. Stuttgart: Metzler.

Goldberg, A. E. (2003). Constructions: a New Theoretical Approach to Language. *TRENDS in Cognitive Sciences*, *7*(5), 219.

Grabe, W. (2009). *Reading in a Second Language. Moving from Theory to Practice*. Cambridge: CUP.

Hacken, P. ten. (2003). Computer-Assisted Language Learning and the Revolution in Computational Linguistics. *Linguistik online*, *17*(05), 23–39.

Hacken, P. ten, Abel, A., & Knapp, J. (2006). Word Formation in an Electronic Learners' Dictionary: ELDIT. *International Journal of Lexicography*, *19*(3), 243.

Hacken, P. ten, & Domenig, M. (1996). Reusable Dictionaries for NLP: The Word Manager Approach. *Lexicology*, *2*, 232–255.

Hacken, P. ten, & Tschichold, C. (2001). Word Manager and CALL: Structured Access to the Lexicon as a Tool for Enriching Learners' Vocabulary. *ReCALL*, *13*(01), 121–131.

Hall, B., & Wan, S. (2002). *Object-Oriented Programming with ActionScript*. New York: Pearson Education.

Hama, M., & Leow, R. P. (2010). Learning without Awarenes Revisited. Extending Williams 2005. *Studies in Second Language Acquisition*, *32*, 465–491.

Harris, R. (1995). *The Linguistics Wars*. New York: OUP.

Heift, T. (1998). *Designed Intelligence: A Language Teacher Model*. Unpublished doctoral dissertation, Simon Fraser University, Burnaby.

Heift, T., & Schulze, M. (2007). *Errors and Intelligence in Computer-Assisted Language Learning: Parsers and Pedagogues*. New York: Routeledge.

Helbig, G., Götze, L., Henrici, G., & Krumm, H.-J. (Eds.). (2001). *Deutsch als Fremdsprache* (Vol. 19). Berlin, New York: de Gruyter.

Hémard, D. (1999). A Methodology for Designing Student-Centred Hypermedia CALL. In R. Debski & M. Levy (Eds.), *WorldCALL: Global Perspectives on Computer-Assisted Language Learining* (pp. 215–228). Lisse: Swets&Zeitlinger.

Higgins, J. (1988). *Language, Learners, and Computers*. London: Longman.

Hock, H. H., & Joseph, B. D. (1996). *Language History, Language Change and Language Relationship. An Introduction to Historical and Comparative Linguistics*. Berlin, New York: de Gruyter.

Hoeksma, J. (1985). *Categorial Morphology*. New York: Garland.

Hopper, P., & Traugott, E. (2003). *Grammaticalization*. Cambridge: CUP.

Hubbard, P. (2006). Evaluating CALL Software. In L. Ducate & N. Arnold (Eds.), *Calling on CALL. From Theory and Research to New Directions in Foreign Language Teaching.* San Marcos, TX: CALICO.

Hubbard, P. (Ed.). (2009). *Computer Assisted Language Learning* (Vol. 1-4). London: Routeledge.

Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge: CUP.

Hunter, J., & Crawford, W. (2001). *Java Servlet Programming*. Sebastopol, CA: O'Reilly Media.

Hutton, G. (2007). *Programming in Haskell*. Cambridge: CUP.

Hymes, D. (1992). The Concept of Communicative Competence Revisited. In M. Putz (Ed.), *Thirty Years of Linguistics Evolution* (p. 31-58). Philadelphia: Benjamins.

Ilarraza, A. de, Maritxalar, A., Maritxalar, M., & Oronoz, M. (1999). IDAZKIDE: An Intelligent Computer-Assisted Language Learning Environment for Second Language Acquisition. *ReCALL*, *11*(Special Issue: Language Processing in CALL), 11–19.

Jackendoff, R. (2003). *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford: OUP.

Jager, S., Nerbonne, J. A., & Essen, A. van (Eds.). (1998). *Language Teaching and Language Technology*. Lisse,NL: Swets&Zeitlinger.

Johns, T. (1991). Should You be Persuaded. Two Samples of Data-Driven Language Materials. *ELR Journal*, *4*(Vol. 4), 1–16.

Jurafsky, D., & Martin, J. (2008). *Speech and Language Processing* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Karttunen, L., & Beesley, K. (2003). *Finite-State Morphology*. Stanford, CA: CSLI.

Katamba, F. (1994). *English Words*. London: Routledge.

Kernighan, B., & Ritchie, D. (1988). *The C Programming Language*. Reading, Massachusetts: Addison-Wesley.

Koskenniemi, K. (1983). *Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production*. Helsinki: University of Helsinki.

Krashen, S. D. (1981). *Second Language Acquisition and Second Language Learning*. Oxford: OUP.

Krashen, S. D. (1982). *Principles and Practice in Second Language Acquisition*. Oxford: Pergamon.

Köster, L. (2001). Wortschatzvermittlung. In G. Helbig, L. Götze, G. Henrici, & H.-J. Krumm (Eds.), *Deutsch als Fremdsprache* (Vol. 19, pp. 887–893). Berlin, New York: de Gruyter.

Larsen-Freeman, D., & Long, M. (1990). *An Introduction to Second Language Acquisition Research*. London: Longman.

Laufer, B. (1992). How much Lexis is Necessary for Reading Comprehension? In P. Arnaud & H. Bejount (Eds.), *Vocabulary and Applied Linguistics.* London: Macmillan.

Laufer, B. (1997). What's in a Word that Makes it Hard or Easy: Some Intralexical Factors that Affect the Learnig of Words. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, Acquisition, and Pedagogy.* Cambridge: CUP.

Lehmann, C. (1995). *Thoughts on Grammaticalization*. München: Lincom Europa.

Leipzig Corpora. (last accessed: 13 September 2010). University of Leipzig. Available from `http://corpora.informatik.uni-leipzig.de/download.html`

Levy, M. (1997). *Computer Assisted Language Learning: Context and Conceptualization*. Oxford: Clarendon Press.

Levy, M. (1999). Design Processes in CALL: Integrating Theory, Research and Evaluation. In K. C. Cameron (Ed.), *Computer Assisted Language Learning (CALL): Media, Design, and Applications* (pp. 83–107). Lisse: Swets&Zeitlinger.

Levy, M. (2001). Scope, Goals and Methods in CALL Research: Questions of Coherence and Autonomy. *ReCALL*, *12*(02), 170–195.

Lezius, W. (2000). Morphy–German Morphology, Part-Of-Speech Tagging and Applications. In *Proceedings of the 9th EURALEX International Congress* (pp. 619–623). Stuttgart: University of Stuttgart.

Lovik, T. A., Guy, J. D., & Chavez, M. (2007). *Vorsprung* (2nd ed.). Boston: Houghton Mifflin.

Lutz, M. (2006). *Programming Python*. Sebastopol, CA: O'Reilly Media, Inc.

Lüdeling, A., Schmid, T., & Kiokpasoglou, S. (2002). On Neoclassical Word Formation in German. In J. Booij Geert & van Marle (Ed.), *Yearbook of Morphology 2001* (pp. 253–283). Dordrecht: Kluwer.

Manning, C., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Boston: MIT Press.

Marchman, V. A., & Thal, D. J. (2005). Words and Grammar. In M. Tomasello & D. I. Slobin (Eds.), *Beyond Nature-Nurture. Essays in Honor of Elizabeth Bates.* Mahwah, New Jersey: Lawrence Erlbaum.

McCarthy, M. (1990). *Vocabulary*. Oxford: OUP.

Meara, P., & Glyn, J. (1987). Vocabulary Size as a Placement Indicator. In P. Grunwell (Ed.), *Applied Linguistics in Society. Papers from the Annual Meeting of the British Association for Applied Linguistics (20th, Nottingham, England, UK).* Nottingham: Nottingham University Press.

Menzel, W., & Schröder, I. (1998). Constraint-Based Diagnosis for Intelligent Language Tutoring Systems. In *Proceedings of the IT&KNOWS Conference at the IFIP '98 Congress, Wien Budabest* (pp. 484–497). Budapest: IT&KNOWS.

Meyer, C. F. (2002). *English Corpus Linguistics : An Introduction*. Cambridge: CUP.

Michaud, L. N., & McCoy, K. F. (2000). Supporting Intelligent Tutoring in CALL by Modeling the User's Grammar. *Proceedings of the 13th Annual International Florida Artificial Intelligence Research Symposium, Special Track on AI in Instructional Software (Orlando, FD)*, *1*, 50–54.

Mitchell, R., & Myles, F. (2004). *Second Language Learning Theories* (2nd ed.). London: Hodder Arnold.

Moeller, J., Adolph, W. R., Hoecherl-Alden, G., Berger, S., & II, J. F. L. (2005). *Deutsch heute* (8th ed.). Boston: Houghton Mifflin.

Morphy. (last accessed: 17 September 2010). Homepage of Wolfgang Lezius. Available from `http://www.wolfganglezius.de/lib/exe/fetch.php?media=cl:mosetup.exe`

Motsch, W. (2004). *Deutsche Wortbildung in Grundzügen*. Berlin: Walter de Gruyter.

MPI - Language Annotation Technologies. (last accessed: 18 September 2010). Max-Planck-Institut für Psycholinguistik. Available from `http://www.lat-mpi.eu/`

Nagata, N. (1992). *A Study of the Effectiveness of Intelligent CALI as an Application of Natural Language Processing*. Unpublished doctoral dissertation, Universtiy of Pittsburgh, Pittsburgh.

Nation, P. (1990). *Teaching and Learning Vocabulary*. New York: Heinle and Heinle.

Nation, P. (1999). Grated Readers and Vocabulary. *Reading in a Foreign Language*, *12*, 355–380.

Nation, P. (2001). *Learning Vocabulary in another Language*. Cambridge: CUP.

Nation, P. (2008). *Teaching Vocabulary. Strategies and Techniques*. Boston: Heinle, Cengage.

Nation, P., & Meara, P. (2002). Vocabulary. In N. Schmitt (Ed.), (pp. 35–54). London: Arnold.

Nerbonne, J. (2003). Computer-Assisted Language Learning and Natural Language Processing. In R. Mitkov (Ed.), *Handbook of Computational Linguistics* (pp. 670–698). Oxford: OUP.

Ogden, C. K. (1930). *Basic English: A General Introduction with Rules and Grammar*. London: Kegan Paul.

Olejarka, A. (2008). *Die Wortbildungsregularitäten des Verbs und ihre Umsetzung in didaktischen Grammatiken für Deusch als Fremdsprache*. München: Iudicium.

Open Office Thesaurus. (last accessed: 13 September 2010). OpenOffice. Available from `http://ftp.services.openoffice.org/pub/OpenOffice.org/contrib/dictionaries/thes_de_DE_v2.zip`

Oxford, R. (1990). *Language Learning Strategies: What Every Teacher Should Know*. Boston: Newbury House.

Oxford, R. (2008). Hero with a Thousand Faces: Learner Autonomy, Learning Strategies and Learning Tactics in Independent Language Learning. In S. Hurd & T. Lewis (Eds.), *Language Learning Strategies in Independent Settings* (pp. 41–66). Bristol: Multilingual Matters.

Pennycook, A. (1997). Cultural Alternatives and Autonomy. In P. Benson & P. Voller (Eds.), *Autonomy and Independence in Language Learning* (pp. 35–53). London: Longman.

Pienemann, M. (1998). *Language Processing and Second Language Development: Processability Theory*. Amsterdam: J. Benjamins.

Pigada, M., & Schmitt, N. (2006). Vocabulary Acquisition from Extensive Reading: A Case Study. *Reading in a Foreign Language*, *18.1*, 1–28.

Plag, I. (1999). *Morphological Productivity: Structural Constraints in English Derivation*. Berlin: Walter de Gruyter.

Pokorny, B. (2009). *Language Frequency Profiling of Written TExts by Students of German as a Foreign Language*. Unpublished master's thesis, Unviersity of Waterloo.

Polenz, P. von. (2000). *Deutsche Sprachgeschichte vom Spätmittelalter bis zur Gegenwart* (2nd ed., Vols. Band 1. Einführung, Grundbegriffe: 14. bis 16. Jahrhundert). Berlin: de Gruyter.

Pollard, C. J. (1988). Categorial Grammar and Phrase Structure Grammar: An Excursion on the Syntax-Sematics Frontier. In R. Oehrle, E. Bach, & D. Wheeler (Eds.), *Categorial Grammars and Natural Language Structures* (pp. 391–416). Dordrecht/-Boston/Lancaster/Tokyo: D. Reidel.

Pollard, C. J., & Sag, I. A. (1987). *Information-Based Syntax and Semantics. Volume 1*. Stanford University: CSLI Publications.

Pollard, C. J., & Sag, I. A. (1994). *Head-Driven Phrase Structure Grammar*. Chicago: University of Chicago Press.

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*, *1*, 1799–1802.

Rall, M. (2001). Grammatikvermittlung. In G. Helbig, L. Götze, G. Henrici, & H.-J. Krumm (Eds.), *Deutsch als Fremdsprache* (Vol. 19, pp. 880–886). Berlin, New York: de Gruyter.

Richards, J. (1976). The Role of Vocabulary Teaching. *TESOL Quarterly*, *10*, 11-89.

Riehemann, S. (1993). *Word Formation in Lexical Type Hierarchies. A Case Study of bar-Adjectives in German*. Unpublished master's thesis, Eberhard-Karls University of Tübingen. Semninar für Sprachwissenschaft.

Riehemann, S. (1998). Type-Based Derivational Morphology. *The Journal of Comparative Germanic Linguistics*, *2*(1), 49–77.

Rings, G. (2001). Wirtschaftskommunikation ohne Komposita und Derivate? Zur Vermittlung von Wortbildungsstrukturen in Theorie und Praxis. *GFL Journal*, *1*, 1–13.

Roark, B., & Sproat, R. (2007). *Computational Approaches to Morphology and Syntax*. Oxford: OUP.

Roosmaa, T., & Prószéky, G. (1998). GLOSSER – Using Language Technology Tools for Reading Texts in a Foreign Language. In S. Jager, J. A. Nerbonne, & A. van Essen (Eds.), *Language Teaching and Language Technology* (pp. 101–107). Lisse,NL: Swets&Zeitlinger.

Römer, C. (2006). *Morphologie des Deutschen*. Tübingen: UTB.

Römer, C., & Matzke, B. (2010). *Der Deutsche Wortschatz. Struktur, Regeln, Merkmale*. Tübingen: Narr.

Sag, I. A., Wasow, T., & Bender, E. M. (2003). *Syntactic Theory: A Formal Introduction* (2nd ed.). Stanford: CSLI Publications.

Scalise, S. (1984). *Generative Morphology*. Dordrecht: Foris.

Schiller GEZ article. (last accessed: 17 September 2010). Spiegel online. Available from `http://www.spiegel.de/kultur/gesellschaft/0,1518,581529,00.html`

Schmenk, B. (2006). CALL, Self-Access and Learner Autonomy: A Linear Process from Heteronomy to Autonomy. In D. K. Theo Harden Bernd Witte (Ed.), *The Concept of Progression in the Teaching and Learning of Foreign Languages* (p. 75-93). Bern: Peter Lang.

Schmidt, R. (1990). The Role of Consciousness in Second Language Learning. *Applied Linguistics*, *11*(2), 129.

Schmidt, R. (1994). Deconstructing Consciousness in Search of a Useful Definition for Applied Linguistics. *AILA Review*, *11*, 11–26.

Schmidt, R. (2001). Attention. In P. Robinson (Ed.), *Cognition and Second Language Instruction.* Cambridge: CUP.

Schmitt, N. (1997). Vocabulary Learning Strategies. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, Acquisition, and Pedagogy.* Cambridge: CUP.

Schmitt, N. (1998). Tracking the Incremental Acquisition of Second Language Vocabulary: A Longitudinal Study. *Language Learning*, *48*, 281–317.

Schmitt, N. (2000). *Vocabulary in Language Learning*. Cambridge: CUP.

Schneider, D., & McCoy, K. F. (1998). Recognizing Syntactic Errors in the Writing of Second Language Learners. *Proceedings of the Thirty-Sixth Annual Meeting of the Association for Computational Linguistics and the Seventeenth International Conference on Computational Linguistics (Montreal)*, *Vol. 2*, 1198–1204.

Schulze, M. (2001). *Textana. Grammar and Grammar Checking in Parser-Based CALL.* Unpublished doctoral dissertation, University of Manchester Institute of Science and Technology, Manchester.

Schulze, M., Wood, P., & Pokorny, B. (submitted). Measuring balanced complexity.

Simmler, F. (1998). *Morphologie des Deutschen*. Berlin: Weidler.

Singleton, D. (1999). *Exploring the Second Language Mental Lexicon*. Cambridge: CUP.

Singleton, D. (2000). *Language and the Lexicon: an Introduction*. London: Arnold.

Stallman, R. M. (2002). *Free Software, Free Society: Selected Essays of Richard M. Stallman*. Boston: GNU Press.

Stanford Tagger. (last accessed: 13 September 2010). Stanford University. Available from `http://nlp.stanford.edu/software/tagger.shtml`

Stanovich, K. (1986). Matthew Effects in Reading: Some Consequences of Individual Differences -in the Acquisition of Literacy. *Reading Research Quarterly*, *21*, 360–407.

Stanovich, K. (2000). *Progress in Understanding Reading. Scientific Foundations and New Frontiers*. New York: Guilford Press.

Stroustrup, B. (1997). *The C++ Programming Language*. Reading, Massachusetts: Addison-Wesley.

Stuttgart Tree Tagger. (last accessed: 13 September 2010). http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/. Available from `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`

SWT. (last accessed: 13 September 2010). The Eclipse project. Available from `http://www.eclipse.org/swt/`

Tomasello, M. (2003). *Constructiong a Language. A Usage-Based Theory of Language Acquisition*. Harvard: Harvard University Press.

Valentin, K. (1978). *Alles von Karl Valentin*. München: Piper.

Wall, L., & Loukides, M. (2000). *Programming Perl*. Sebastopol, CA: O'Reilly.

Weinberg, A., Garman, J., Martin, J., & Merlo, P. (1995). A Principle-Based Parser for Foreign Language Training in German and Arabic. In *Intelligent Language Tutors: Theory Shaping Technology* (pp. 23–44). Hillsdale, NJ: Lawrence Erlbaum.

White, C. (2008). Language Learning Strategies in Independent Language Learning. Strategies and Tactics in Independent Language Learning. In S. Hurd & T. Lewis

(Eds.), *Language Learning Strategies in Independent Settings* (pp. 3–24). Bristol: Multilingual Matters.

Williams, J. N. (2005). Learning Without Awarenes. *Studies in Second Language Acquisition*, *27*, 269–304.

Wood, P. (2002). *Zur Produktivität von Derivationsmorphemen: Die Adjektivbildung im Deutschen*. Unpublished master's thesis, University of Siegen, Siegen, Germany.

Wood, P. (2007). *Teaching Vocabulary and Word Formation.* (Paper presented at the ACLA conference 2007 in Saskatoon)

Wood, P. (2008). Developing ICALL Tools Using GATE. *Computer Assisted Language Learning*, *21:4*, 383-392.

Wood, P. (2010). Transparent Language System Complete Edition (German). Software Review. *Calico Journal*, *28.1*, 229-237.

Wood, P. (accepted for publication in 2011). Computer Assisted Reading in German as a Foreign Language. Developing and Testing a NLP Based Application. *Calico Journal*, *Special Issue: CALL in Canada*.

Wood, P. (submitted). QuickAssist: Reading and Learning Vocabulary Independently With the Help of CALL and NLP Technologies.

Zimmermann, C. B. (1997). Second Language Vocabulary Acquisition. In J. Coady & T. Huckin (Eds.), (pp. 5–19). Cambridge: CUP.