Volume 26 - Number 1 - 2011 - ISSN 2190-6858

JLCL
Journal for Language Technology
and Computational Linguistics

Language Resources and Technologies in Learning and Teaching

Sprachressourcen und -technologien in Lehre und Lernen

Herausgegeben von */ Edited by*
Maja Bärenfänger, Frank Binder, Henning Lobin, Harald Lüngen, Maik Stührenberg

GSCL Gesellschaft für Sprachtechnologie & Computerlinguistik

Contents

Impressum

Herausgeber	Gesellschaft für Sprachtechnologie und Computerlinguistik (GSCL)
Aktuelle Ausgabe	Band 26 – 2011 – Heft 1 "Language Resources and Technologies in Learning and Teaching"
Gastherausgeber	Maja Bärenfänger, Frank Binder, Henning Lobin, Harald Lüngen, Maik Stührenberg
Anschrift der Redaktion	Lothar Lemnitzer Berlin-Brandenburgische Akademie der Wissenschaften Jägerstr. 22/23 10117 Berlin lemnitzer@bbaw.de
ISSN	2190-6858
Erscheinungsweise	2 Hefte im Jahr, Publikation nur elektronisch
Online-Präsenz	www.jlcl.org

Maja Bärenfänger, Frank Binder, Henning Lobin, Harald Lüngen, Maik Stührenberg

Editorial

This volume contains selected and revised papers presented at two workshops focusing on different roles that language resources & technologies may have in learning & teaching. Part one of this special issue is entitled "E-Learning and Computational Linguistics" and includes three papers that were presented at the KONVENS 2010 post-conference workshop "Language Technology and Text Technological Methods for E-Learning" (September 8th, Saarbrücken). An overview of these three papers and the overall research field "E-Learning and Computational Linguistics" is given in Bärenfänger/Stührenberg (in this issue). The second part of this volume is entitled "Sprachressourcen in der Lehre" ("Language Resources in Academic Training and Teaching") and emerged from a D-SPIN WP6 workshop by the same title, which was held on January 18th 2011 at the Berlin-Brandenburg Academy of Sciences and Humanities in Berlin. Binder/Lüngen/Lobin (in this issue) provide an introduction to that section, followed by five contributions from the workshop.

The guest-editors would like to thank the editor-in-chief, Lothar Lemnitzer, for his support and for giving us the opportunity to publish these papers in this special issue of the JLCL. We would also like to thank Anna Melzer for formatting the articles of this issue.

Der vorliegende Band enthält ausgewählte und überarbeitete Beiträge zweier Workshops, welche sich mit verschiedenen Rollen von Sprachressourcen und -technologien in Lehre und Lernen beschäftigten. Der erste Teil dieser Sonderausgabe trägt den Titel „E-Learning and Computational Linguistics" und beinhaltet drei Artikel, die aus Präsentationen des Workshops „Language Technology and Text Technological Methods for E-Learning" im Rahmen der KONVENS 2010 am 8. September 2010 hervorgegangen sind. Einen Überblick darüber und die Einordnung dieser Arbeiten in das Forschungsfeld „E-Learning und Computerlinguistik" ist im Beitrag von Bärenfänger und Stührenberg zu finden. Der zweite Teil des Heftes trägt den Titel „Sprachressourcen in der Lehre" und präsentiert Artikel des gleichnamigen D-SPIN-AP6-Workshops, der am 18. Januar 2011 an der Berlin-Brandenburgischen Akademie der Wissenschaften in Berlin abgehalten wurde. Der Beitrag von Binder, Lüngen und Lobin liefert den Überblick zu diesem Teil des Heftes mit seinen fünf Beiträgen.

Als Gast-Herausgeber möchten wir unseren Dank gegenüber dem Chefredakteur Lothar Lemnitzer ausdrücken, der uns die Möglichkeit gegeben hat, diese Artikel im vorliegenden JLCL-Sonderband zu veröffentlichen. Darüber hinaus gilt unser Dank Anna Melzer für den Satz dieser Ausgabe.

Maja Bärenfänger, Maik Stührenberg

E-Learning and Computational Linguistics An Introduction

0 Preface

E-Learning is more and more becoming a commodity in all areas of education ranging from curricular learning at schools and universities via further education and learning on the job to social and game-based learning in spare time activities. At the same time, research on E-Learning is scattered across different research disciplines, being often a matter of single initiatives and persons. In the light of this growing gap, the authors of this introduction want to showcase Computational Linguistics and Language Technology as possible key enabling technologies for current E-Learning.

1 Current trends in E-Learning

E-Learning, or Technology-Enhanced Learning is an interdisciplinary research area which involves research disciplines like Pedagogy, Pedagogical Psychology, Computer Science (particularly research on human-computer interaction (HCI), knowledge management, and software architectures for virtual learning management systems), Artificial Intelligence (especially research on intelligent tutoring systems (ITS)) and Computational Linguistics (in particular computer-assisted language learning, educational natural language processing and text technology for learning technology standards and applications). In the last 15 years, the research area "E-Learning" has undergone major changes. While the first decade, starting in the end of the 20th century, has mainly focused on learning infrastructure and learning content, more recently, the learning process and the learners have been taken more seriously. Today, a learner-centred paradigm dominates E-Learning. Current areas of interest are adaptivity and personalization (individual and ubiquitous learning settings, personal infrastructures supporting life long learning), learning in social communities, and learning & entertainment (game-based learning).

Taking up trends in the Web 2.0, attention has been given to aspects like virtual learning communities, web based collaboration, social media content and social learning infrastructures (e.g. edublogs and wikis). This insight also had an influence in research on and development of virtual learning environments: Personal learning environments or mashup personal learning environments are subjects of increasing interest, and are more and more seen as more adequate promoters of life long learning than traditional virtual learning environments. Moreover, applications and methods supporting learning at any time and any place are also seen as important factors to improve life long learning. Catch phrases in this context are "ubiquitous learning", "mobile learning" and "augmented reality". It can be expected that these research and development fields will become more and more important in the next years. Beside these trends, two more important aspects of current technology-enhanced learning have to be mentioned: E-assessment and game-based learning. E-Assessment has always played a major role in E-Learning, but has long been restricted to simple multiple-

choice or fill-in-the-gap tasks, because of a lack of more advanced methods to automatically evaluate input from open tasks or essays. In the last years, there has been profound progress in this area, especially through the application of educational natural language processing techniques. Game-based learning or serious games are the flip side of the coin. These areas focus on more informal and entertaining methods of training and self-assessment.

2 E-Learning and Computational Linguistics

Computational linguistics (CL) has a long tradition in research on E-Learning. For a long time, this research has been restricted to computer-assisted language learning (CALL). But in the last decade, attention has been shifted to other research areas concerning a large variety of aspects of E-Learning. In the following, we will give an overview of main areas and aspects. This overview is organized in three parts: 1. CL research on digital learning content and resources, 2. CL research on computer-mediated communication and collaboration, and 3. CL research on E-assessment.

1. **Digital learning content and resources**: Computational linguistics research on content is concerned with the following major tasks: a) the development and annotation of language resources for E-Learning (corpora, ontologies, and semantic resources), b) the analysis and evaluation of content (collaboratively constructed resources/user-generated content vs. learning objects). The first task is especially related to Text Technology. Text Technological knowledge is applied when it comes to structuring content in a modular way. Separating content and formatting by using methods such as XML allows for E-Learning content to be presented in various formats and screen sizes, establishing the foundation for mobile and/or ubiquitous learning. The second task deals with a variety of aspects that range from information extraction (e.g. term, keyword, glossary or definition extraction), information visualization (e.g. automatic structure discovery and visualization, concept visualization) to quality assessment (e.g. opinion mining, sentiment analysis).

2. **Computer-mediated communication and collaboration:** Computational linguistics research in this area is primarily concerned with the development of tools that support collaborative work (especially tasks like editing, searching, evaluating – e.g. by information/structure/concept visualization, grammatical error detection and correction, glossary and definition detection, and quality assessment) and communication between learners (e.g. by summarizing discussion threads) or between learner and virtual learning environment (e.g. by providing an E-tutor/generating tutorial responses). The latter can be viewed as a special case of an adaptive dialogue system.

3. **E-assessment:** In the context of E-assessment, three types of computational linguistics research tasks can be distinguished: a) analysis of learner input (in particular grammatical error detection and correction, discourse and stylistic analysis, plagiarism detection), b) generation of feedback and generation of test question/tests, and c) monitoring learning process (e.g. visualization of learner's concept maps).

In the mentioned research fields, a variety of methods and techniques are used. These methods include, but are not limited to educational natural language processing techniques

(Educational NLP), or methods referring to text or data mining, latent semantic analysis (LSA), information retrieval and extraction.

As a result, E-Learning can be seen as a prime application of Computational Linguistics and Text Technology, both as a field in which knowledge and methods can be applied in an integrated manner but also as a research topic. In both ways, CL and its sub-disciplines can act as hinge for other disciplines involved such as Pedagogy or Computer Science. Still, many technologies are in their infancies and thus lack the robustness and usability needed for professional product development. At the same time, the users of E-Learning are typically open to new technologies and a playful use even of error prone beta versions of software.

One problem that has been observed from experts in the area is a difficulty in getting funding for interdisciplinary research in general and E-Learning-related projects in particular. Current subject-internal reviewing processes often seem to be an obstacle in launching new projects. Substantial funding and longer lasting programs would definitely help establishing Language Technology as an integral part of E-Learning software. The market opportunities for education technology, especially in game-based learning are huge. The authors of this paper hope to help creating more acceptance for the research to be done on this topic.

3 Important Conferences and Workshops

E-Learning as a highly interdisciplinary research and development field deals with a wide variety of aspects on learning and technology. Related conferences and workshops therefore often offer a broad spectrum of topics from different disciplines. Important broadly oriented E-Learning conferences are:

- CSEDU (International Conference on Computer Supported Education, since 2009)
- EC-TEL (European Conference on Technology Enhanced Learning, since 2006)
- ECEL (European Conference on e-Learning, since 2002)
- ONLINE EDUCA (International Conference on Technology Supported Learning & Training)
- ICL (International Conference on Interactive Computer aided Learning, since 1998)
- IEEE EDUCON conference (since 2010)

On the other hand, E-Learning is also discussed in more special contexts, e.g. in communities which focus on one specific aspect of E-Learning. Regular conferences and workshops that are especially relevant for computational linguists dealing with E-Learning are:

- ACL-HLT Workshop on Innovative Use of NLP for Building Educational Applications (since 2003)
- DeLFI (e-Learning-Fachtagung Informatik der Gesellschaft für Informatik, since 2003)

- LREC (International Conference on Language Resources and Evaluation, since 1998)
- RANLP (Recent Advances in Natural Language Processing, since 1995)

These conferences often offer workshops on "E-Learning and Computational Linguistics/Language Technology/Natural Language Processing". In the last four years, workshops had discussed topics like

- "The People's Web Meets NLP: Collaboratively Constructed Semantic Resources" (workshop in conjunction with ACL 2009 and workshop in conjunction with COLING 2010),
- "Supporting eLearning with Language Resources and Semantic Data" (workshop in conjunction with LREC 2010),
- "Natural Language Processing and Knowledge Representation for eLearning environments", "NLP for Educational Resources" (workshops in conjunction with RANLP 2007),
- "What can Natural Language Processing and Semantic Web technologies do for eLearning?" (workshop in conjunction with ACL 2007).

In contrast to the above-mentioned workshops and conferences that are open for a wide variety of topics in the context of "E-Learning and Computational Linguistics/Language Technology/Natural Language Processing", the following conferences focus on a single special research and development field, namely computer-assisted language learning (CALL). Research and development on CALL has long been the only computational linguistics research area in E-Learning. Conferences in this field often have a long tradition and go back to the beginning of the 1990s. Important conferences on this topic are:

- EUROCALL (Conference of the European Association for Computer Assisted Language Learning, since 1993)
- WORLDCALL (Conference of the Worldwide Association for Teachers and Educators interested in Computer Assisted Language Learning, since 1998)
- CALICO (Conference of the Computer Assisted Language Instruction Consortium, since 1998)
- IALLT (Conference of the International Association for Language Learning Technology, since 2000)

These lists are not exhaustive and are only included to show both the broad spectrum of E-Learning as a research topic as such and the significance that E-Learning has gained in the CL community.

4 Special interest group "Language technology and text technological methods in E-Learning" (GSCL-SIG "E-Learning")

In 2007, "Language technology and text technological methods in E-Learning" was constituted as a special interest group of the German Society for Computational Linguistics and Language Technology (GSCL). The aim of this special interest group is to shape the Computational Linguistics and Text Technological perspective on E-Learning, particularly on aspects like personalization and adaptivity, and to elaborate on and discuss methods and applications that can be assigned to E-Learning specific tasks. The special interest group serves as a panel to support communication and cooperation between experts with different research profiles and competencies (hypermedia, natural language processing, learning technology standards, computer-assisted language learning, amongst others).

As a step towards an opening of the discussion to a wider audience, the GSCL-SIG organized a workshop on "Language Technology and Text Technological Methods for E-Learning" which took place on the 8th September 2010, in conjunction with KONVENS 2010. The workshop presented five plenary talks on various aspects of the workshop topic and two invited keynote talks on "Wikulu: Information Management in Wikis Enhanced by Language Technologies" (by Iryna Gurevych, TU Darmstadt) and "The Snowflake Effect in learning and research" (by Erik Duval, Katholieke Universiteit Leuven). The complete program of the workshop is available at "http://konvens2010.coli.uni-saarland.de /wsprogram_en.html".

5 Overview of papers in the first part of this volume

The first part of this special issue contains three papers that emerged from presentations at the workshop. The authors considerably extended and revised their original submissions so that the articles in this volume represent the current state of the art of the authors' work (see table 1 for an overview).[1]

The contribution "Using Latent-Semantic Analysis and Network Analysis for Monitoring Conceptual Development" by Fridolin Wild, Debra Haley and Katja Bülow focuses on language technology for monitoring learning progress. The authors report on a system called CONSPECT which helps online learners and tutors to monitor the learner's conceptual development by extracting and visualising a conceptual representation from a learner's blog post or the like. For the concept extraction, Latent Semantic Analysis (LSA) and Network Analysis (NA) are combined to a method called Meaningful Interaction Analysis (MIA). The system is evaluated in two verification experiments. In the first experiment, the output of CONSPECT (namely, concept clusters) is compared to human output (namely, human card sorts) in order to find out how closely humans agree with the system's concept clusters. The second experiment is used to evaluate if human text annotations are similar/different to text annotations made by CONSPECT. Wild/Haley & Bülow appraise the findings of both

[1] The papers in this part of the special issue were peer-reviewed in two rounds (extended abstract and full paper) by at least two members of the program committee of the workshop and by the editors of part one, Maja Bärenfänger and Maik Stührenberg. Through this review process, three out of seven submitted papers were selected for this publication.

experiments as a promising start for a computer-based system that monitors conceptual development.

In their article "Meaning versus Form in Computer-assisted Task-based Language Learning: A Case Study on the German Dative", Sabrina Wilske & Magdalena Wolska report on an empirical study which investigated the learning effects of different types of production settings (free vs. constrained production) and different types of feedback (implicit vs. explicit/metalinguistic) in a dialogue system. This system supports a goal-oriented communicative approach to language learning for German. In this approach, communication between learner and dialogue system is framed in a real world situation – in this case triggered by a directions giving task – that elicit the use of the dative case in prepositional phrases. Advantages of this approach are a clearly defined communicative outcome and a focus on meaning. The main focus of this article lies on the empirically grounded comparison of learning effects between free vs. constrained production and implicit vs. explicit feedback. The results of the empirical study indicate that a stronger focus on form (constrained production) has greater learning effects, especially with regard to accuracy gains in using the form.

	Wild/Haley/Bülow	**Wilske/Wolska**	**Miyakoda/Kaneko/ Ishikawa**
Research Area	E-assessment/(self-) awareness of the learner's progress, particularly by computer-based analyses of states in a learner's conceptual development	Computer-mediated communication, in particular task-based instructional dialogues with a computer-based language learning system	Digital learning content, namely multimodal user-generated items for vocabulary learning
Learning Setting	Self-directed learning of factual knowledge in a personal learning environment	Instructed learning of grammatical structures (namely: the dative case in prepositional phrases) in a computer-based dialogue system	Self-directed vocabulary-learning in a personal and ubiquitious learning environment for mobile devices
Methods	Latent Semantic Analysis and Network Analysis	NLP methods (e.g. parsing with mal-rules, template-based dialogue generation)	Text technological methods (in particular: mobile web applications)
Domain	Factual knowledge learning	Language learning	Language learning

Table 1: Overview of the papers

The contribution "Effective Learning Material for Mobile Devices: Visual Data vs. Aural Data vs. Text Data" (Haruko Miyakoda, Kei-ichi Kaneko & Masatoshi Ishikawa) introduces

a vocabulary learning system for mobile devices which supports learners in creating their own multilingual and multimodal learning material, distributing and sharing it with other learners as a podcast via the iTunes application, and evaluating their own material or material made by others by score or comment. Like Wilske/Wolska, the authors also investigate the dependency between learning effect and learning setting – their focus lies on the effect of different types of vocabulary material (text, visual, aural, and combinations of these) on memory retention rate/memorization. For this purpose, two experiments with 100 learners of German as a foreign language are carried out that show that – contrary to former claims – visual data does not provide much aid in vocabulary memorization.

The three papers are as diverse as the research field that has been outlined in Section 2, adding pieces to the E-Learning mosaic: They touch different research areas, are related to different learning settings, and use different methods to achieve their respective research aim:

All contributions did carry out empirical studies and tests – either to evaluate the implemented system in comparison to human performances, or to evaluate which one of several alternatives – e.g. which type of learning content (visual, audio, text) or which type of feedback (focus on form vs. focus on meaning) – induces better learning effects. While the contribution by Wild/Haley/Bülow primarily focuses on technical and methodical aspects, Wilske/Wolska and Miyakoda/Kaneko/Ishikawa especially concentrate on empirical learner studies and the dependency between learning effects and learning setting.

In general these three papers should convince the reader of the broad distribution of Computational Linguistics approaches in the field of E-Learning.

6 Acknowledgments

This volume would not have been possible without the help of many people. First of all, we would like to thank the editor-in-chief of the JLCL, Lothar Lemnitzer, for his support and patience in the edition process. Our special thanks go to Aljoscha Burchardt for his valuable contribution to this introductory article: The preface and some thoughts that lead to the statements in Section 2 go back to him. We also would like to thank the program committee for their reviews of the submitted papers – not also for this special issue, but also for the preceding workshop: Delphine Bernhard, Lothar Lemnitzer, Henning Lobin, Cerstin Mahlow, Michael Piotrowski, Csilla Puskás, Tonio Wandmacher, Eline Westerhout, Fridolin Wild. The final responsibility for accepting or rejecting submissions or for requesting a second or third revision lies with Maja Bärenfänger and Maik Stührenberg.

We would also like to thank the Organizing Committee of KONVENS 2011 and the department of Computational Linguistics & Phonetics at Saarland University for making this workshop become possible. And of course, we want to express our thanks to all speakers who presented their research on our workshop, and to the authors for compiling paper versions of their talks. Last but not least, we would like to thank all participants of the workshop on “Language Technology and Text Technological Methods for E-Learning”, that not only made the workshop very successful and interesting but also laid the foundation for this special issue.

Fridolin Wild, Debra Haley, Katja Bülow

Using Latent-Semantic Analysis and Network Analysis for Monitoring Conceptual Development

This paper describes and evaluates CONSPECT (from concept inspection), an application that analyses states in a learner's conceptual development. It was designed to help online learners and their tutors monitor conceptual development and also to help reduce the workload of tutors monitoring a learner's conceptual development.

CONSPECT combines two technologies - Latent Semantic Analysis (LSA) and Network Analysis (NA) into a technique called Meaningful Interaction Analysis (MIA). LSA analyses the meaning in the textual digital traces left behind by learners in their learning journey; NA provides the analytic instrument to investigate (visually) the semantic structures identified by LSA.

This paper describes the validation activities undertaken to show how well LSA matches first year medical students in 1) grouping similar concepts and 2) annotating text.

1 Theoretical Justification

This section mentions two related Cognitive Linguistic theories that support the approach taken in CONSPECT: Fauconnier's Mental Spaces Theory and Conceptual Blending Theory (Evans and Green 2006). These theories hold that the meaning of a sentence cannot be determined without considering the context. Meaning construction results from the development of mental spaces, also known as conceptual structures (Saeed 2009), and the mapping between these spaces.

Mental spaces and their relationships are what LSA tries to quantify. LSA uses words in their contexts to calculate associative closeness among terms, among documents, and among terms and documents. This use of context is consistent with Fauconnier's claim that context is crucial to construct meaning.

Various researchers use network analysis to analyse conceptual structures: Schvaneveldt et al (1989), Goldsmith et al (1991) and Clariana & Wallace (2007) are among the researchers who use a particular class of networks called Pathfinder, which are derived from proximity data (Schvaneveldt, Durso et al. 1989). These researchers assume that "concepts and their relationships can be represented by a structure consisting of nodes (concepts) and links (relations)." The strength of the relationships can be measured by the link weights. The networks of novices and experts are compared to gauge the learning of the novices.

Pathfinder techniques require the creation of proximity matrices by association, or relationship testing. LSA, on the other hand, requires no such explicit proximity judgments. It uses textual passages to compute automatically a proximity matrix. Thus LSA requires less human effort than these other techniques.

1.1 Latent Semantic Analysis

The subsection briefly explains LSA, a statistical natural language processing technique whose purpose is to analyse text. For a comprehensive introduction to LSA, see Landauer et al (2007). LSA was chosen as the analysis technique due to the vast literature reporting positive results and to the authors' extensive research into LSA (Haley 2008; Wild 2010). Other tools exist but have not been tested extensively in an educational setting. In addition, they are not grounded in a cognitive framework: they are mostly the results of improvements at the level of basic similarity measures.

LSA is similar to the vector space model (Salton, Wong et al. 1975), which uses a large corpus related to the knowledge domain of interest and creates a term/document matrix whose entries are the number of times each term appears in each document. The LSA innovation is to transform the matrix using singular value decomposition (SVD) and reduce the number of dimensions of the singular value matrix produced by SVD, thus reducing noise due to chance and idiosyncratic word choice. The result provides information about the concepts in the documents as well as numbers that reflect associative closeness between terms and documents, terms and terms, and documents and documents.

2 Technology Description

2.1 Overview

CONSPECT, a web-based, widgetised service accepts RSS feeds as input, processes the data using LSA and network analysis, and outputs results non-graphically as lists or visually in the form of conceptograms. Conceptograms are a type of visualisation conceived and developed by the first author. See (Wild, Haley et al. 2010) for more information.

The RSS feeds are in the form of text from blog posts or learning diaries. These posts and diaries are assumed to be a normal part of a learner's course work or a tutor's preparation. It would be possible to write the code to allow Word documents as input; however, time constraints have not allowed this enhancement. The decision was made to use feeds from blogs rather than Word documents because learner reflection in the form of blogs or online learning diaries has become common.

CONSPECT allows the user to include all or some of the blog posts in the feed, thus providing more flexibility. An example of a conceptogram is shown in Figure 1 and is described below.

2.2 The user point of view

After logging in using openID, the learner is shown a list of existing RSS feeds and conceptual graphs, called conceptograms. Each graph can be inspected in a visualisation using force-directed layouts (conceptograms) of the output of the LSA processing. The user can add a new feed, view a conceptogram, or create an agreement plot between two distinct conceptual representations, i.e., a combined conceptogram.

A single conceptogram shows the concepts written about in the feed. Various types of single conceptograms can be produced using appropriate feeds. An individual student conceptogram would show which concepts the student has written about in the blog. A single conceptogram showing the course's intended learning outcomes (ILO) would come from a tutor-provided blog post giving the ILO. A combined conceptogram compares the concepts of two graphs; for example, if the learner compares a conceptogram showing a course's intended learning outcomes with the conceptogram of his personal learning history, he can see which of the intended outcomes he has covered, which he has not covered, and which concepts he has written about that go beyond the intended learning outcomes.

Similarly, a tutor can monitor the progress of her learners. By aggregating conceptual graphs of several learners, reference models can be constructed. Other possibilities are to compare one learner's conceptograms over time and to compare a learner's conceptogram to the group's emergent reference model (created by combining individual student conceptograms covering a particular time frame). Figure 1 shows a combined conceptogram that compares the concepts of two learners. The real version has three colours - one colour would show the concepts discussed by both students, one colour would show concepts discussed by Student 1 but not Student 2 and similarly for the third colour. Figure 1 shows that both students discussed *type*, *diabet*, *insulin*, and *time*. (The words are stemmed.) Student 1 wrote about *experi*, *parent*, *child*, and *matern* among other concepts, none of which Student 2 covered. Some of the concepts that Student 2 wrote about that were neglected by Student 1 were *nsaid*, *treatment*, *heart*, *obes*, and *glucos*. The term, *nsaid*, is an interesting example of the importance of the existing knowledge of a participant. In this experiment, the participants were medical students and would thus be expected to know that *nsaid* stands for non-steroidal anti-inflammatory drug.

2.3 The background processing

A great deal of processing takes place before the user can see a conceptogram. First, an LSA semantic space must be created from a knowledge domain-specific training corpus. (For the experiments in this paper, the corpus comprised 24,346 Pubmed documents resulting in 21,091 terms. The input was stemmed. Three hundred dimensions were used to reduce the middle matrix of the singular value decomposition.) Next, a feed is used like a typical LSA document; it is converted to and folded in to the original semantic space, i.e., the space created by the original LSA processing. The concepts are filtered so that those closeness relations with very high cosine proximities below the

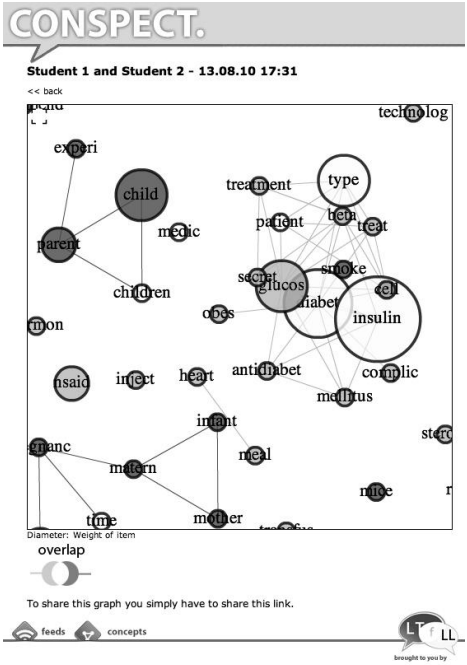


Figure 1: A combined conceptogram showing, overlapping and non-overlapping concepts converted to gray scale for publishing

similarity threshold of 0.7 are assigned zero and eliminated from the latent semantic network, thereby filtering for the most central concepts of the texts investigated. Next, ideas from network analysis (Brandes and Erlebach 2005) are used to identify structural properties of the graph. Several rendering techniques re-represent the conceptual graph structure to the end-user. Finally, the graphs are displayed using a force-directed layout technique (Fruchterman and Reingold 1991) which is deployed to create 2D representations.

3 Evaluation: Verification

Two types of evaluation of CONSPECT were carried out: verification and validation. Verification tries to determine if the system was built correctly while validation looks at whether the right system was built. The validation results are discussed elsewhere. This section describes the two verification experiments that were conducted: cluster analysis and text annotation. Eighteen first year medical students participated in both

experiments; by chance, half were female and half were male. They ranged in age from about eighteen to twenty-two. Each student received a £10 book voucher for participating. Being medical students, the students could be expected to be familiar with terms in the experiment.

3.1 Experiment 1: clustering

The accuracy of CONSPECT was verified in Experiment 1, which examined whether humans cluster concepts in the same way as does CONSPECT. It was a type of card-sorting (Rugg and McGeorge 1997), a technique often used by web designers but used here in a more unusual way. Card sorts allow a researcher to view a participant's mental model of the words on the cards, which is exactly what was wanted. There is a rich literature on how to conduct card sorts (Rugg and McGeorge 1997; Upchurch, Rugg et al. 2001) particularly relating to web page design, which is characterised by a relatively small number of words. This kind of data is often interpreted qualitatively. It is harder to find advice on how to interpret card sorts with a large number of words. Deibel (2005) encountered just such a problem and developed the concept of edit distance of card sorts to analyze her data. The edit distance is the number of cards from one card sort that must be moved from one pile to another in order to match another card sort.

3.1.1 Methodology

Preparation: CONSPECT generated a list of about 50 concepts for five documents from authentic postings about "safe prescribing". (These concepts were chosen randomly; the LSA cosine similarity measures were used to group concepts into categories.) The concepts were printed on a set of cards; this yielded five sets of about 50 cards in each set for each participant.

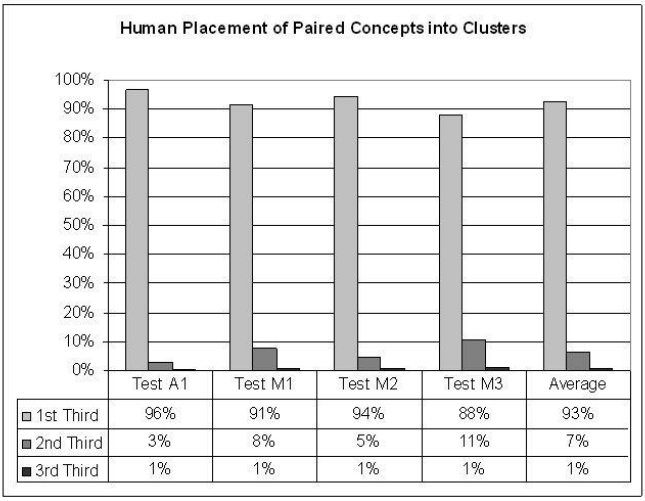
Procedure: The researcher gave sets of cards to the participants and asked them to arrange the cards into groups so that each group contained semantically similar concepts. The participants decided on the number of categories but it had to be more than one and less than the number of cards in the set, that is, there had to be more than one category and each category had to have more than one card. The experimenter then recorded the concepts and the categories chosen by the participant.

3.1.2 Discussion

The analysis provided information on how closely humans agree with CONSPECT's concept classifications. (The classes arise from the LSA cosine similarity measures.) This analysis was undertaken in three ways.

First, the researcher used co-occurrence matrices. Figure 2 shows the spread of data from the co-occurrence matrices. The bar chart shows a noted similarity between the four postings. On average, the vast majority of the paired concepts were in the bottom third, that is, 93% of the pairs were put in the same group by from 0 to 6 participants. Just 7% of the pairs had between 7 and 12 participants placing them in the same cluster.

A tiny number, just 1% of the pairs, were placed in the same cluster by more than 12 of the participants. These groups are referred to as the first, second, and third "thirds".



	Test A1	Test M1	Test M2	Test M3	Average
1st Third	96%	91%	94%	88%	93%
2nd Third	3%	8%	5%	11%	7%
3rd Third	1%	1%	1%	1%	1%

Figure 2: Human Placement of Concepts in Clusters

The second analysis used Deibel et al's (2005) metric of edit distances. The analysis showed that the 18 human participants were about 10% better than was CONSPECT in clustering concepts. Table 1 shows the results of four card sorts, each sort conducted by 18 participants. The table reports on the minimum, maximum, and 1st, 2nd, and 3rd quartile edit distances found by the UW Card Sort Analyzer [2010] for the participants. The lines labelled CONSPECT show the same information when it was compared with the 18 participants. (CONSPECT's sorting data are the clusters calculated by LSA.)

By looking at the edit distance information, one can compare how CONSPECT performs in relation to the human participants. For the min, max, and average quartile edit distances, the CONSPECT figures are larger in each case.

Table 2 shows the results of an attempt to further understand the card sorting data from Table 1. An interesting question was whether or not the edit distances were dependent on a particular variable. The first column, *Sort*, is the number of the sort. The second column, *difference*, was calculated by subtracting the average of the means. The third column, *%diff*, is the difference divided by average of the means. The fourth column, *#cards*, is the number of cards sorted by the participants. The fifth column, *%of cards* is the *#cards* divided by the average of the means, so this indicates, for example, that out of a total of 52 cards for Sort 3, 67% of them had to be moved (i.e., the edit distance) to achieve identical sort piles. Finally, the last column, *words in posts*, is the number of words captured in the blog and used to extract the concepts to

		Card Sorting Results in Terms of Edit Distance					
		Min	Quart. 1	Quart. 2	Quart. 3	Max	#comparisons
Test A1	Avg.	21	26.6	28.7	29.8	34	153
MIA		30	31.25	33	34	35	18
Test M1	Avg.	18	22.6	24.7	26.4	31	153
MIA		24	27	28	29	31	18
Test M2	Avg.	21	30.9	32.7	34.5	40	153
MIA		31	33.3	35.5	36	37	18
Test M3	Avg.	20	28.7	31.0	33.1	38	153
MIA		30	31	32.5	35	37	18

Table 1: Card Sort Result

be sorted. There is no clear relationship among these variables. Therefore, one cannot say that shorter posts result in larger edit distances, for example.

The third column indicates how much larger (as a percentage) the edit distances

sort	diff	%diff	#cards	%cards	#words
3	2.26	6.5%	52	67%	1117
4	2.33	7.0%	51	65%	533
2	3.2	11.5%	43	65%	557
1	4.5	13.6%	48	68%	228
average	9.7%				

Table 2: Edit Distance Information

were for CONSPECT than for the human participants. These figures range from 6.5% to 13.6% with a mean of 9.7%. This analysis suggests that CONSPECT has an edit distance of about 10% larger than the human participants.

The third type of analysis created silhouette plots that showed how well CONSPECT created its clusters. Figure 3 shows the plots. The average silhouette width is .09 for CONSPECT and between -.1 and -.03 for the participants. This means that although the machine was clustering slightly better than the participants, the clusters chosen in all 19 cases were not necessarily very discriminate (but also definitely not bad, which would have been reflected in an average silhouette width of -1.0).

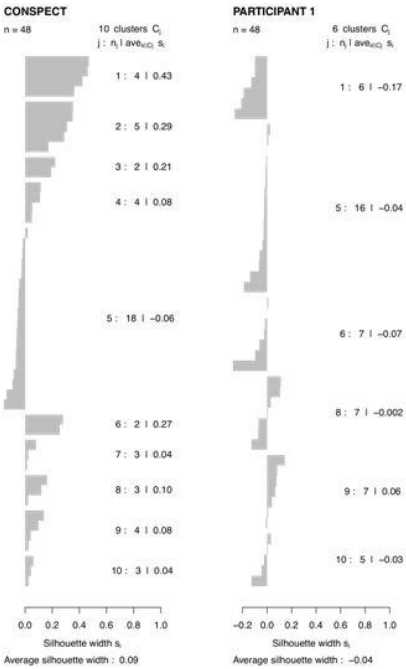


Figure 3: Silhouette plots

3.2 Experiment 2: text annotation

Experiment 2 looked at whether humans agreed with the descriptors that CONSPECT assigned to a text, i.e., it compared the annotations that humans made to a text with CONSPECT's annotations. The same participants were used as were used in the card sorting experiment.

3.2.1 Methodology

Preparation: CONSPECT generated ten descriptors (those with the highest similarity) for each of five texts obtained from postings about safe prescribing (selected randomly from authentic postings from students following a medical course); additionally, five "distracters" were chosen randomly from the available vocabulary. These fifteen descriptors were printed in alphabetical order on a sheet of paper along with the text of the posting.

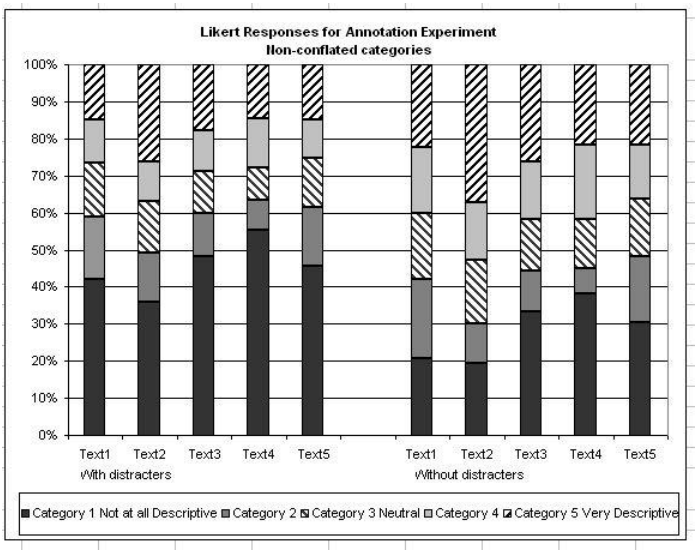


Figure 4: Non-Conflated Likert Responses for Annotation Exper

Procedure: Each participant was given five sheets of paper, one for each test, and were asked to rank each descriptor on a Likert scale of 1 to 5 based on whether they thought the concept was descriptive of the post.

3.2.2 Discussion

Three techniques were used to analyse the text annotation data. The first and second techniques to analyse the text annotation data used the free marginal kappa figure (Randolph 2005; Randolph 2005) a type of inter-rater reliability statistic that is applicable when the raters are not constrained by the number of entries per category. The data come from the Likert selections, that is, the judgments of the participants as to how closely a concept described a text.

Figure 4 and Figure 5, which show stacked bar charts for non-conflated and conflated categories, respectively. From the bottom, the Likert categories were "not at all descriptive", "not very descriptive", "neutral", "somewhat descriptive" and "very descriptive". When distracters are used, more descriptors fall into the bottom two categories - not surprising since distracters were randomly selected and not chosen for their high similarity to the text. Figure 5 is a bit easier to interpret - the two bottom categories were conflated, as were the two top categories.

Tables 3 and 4 below show a different type of analysis. Table 3 shows the results for five categories; Table 4 shows the results for 3 categories (i.e., categories 1 and

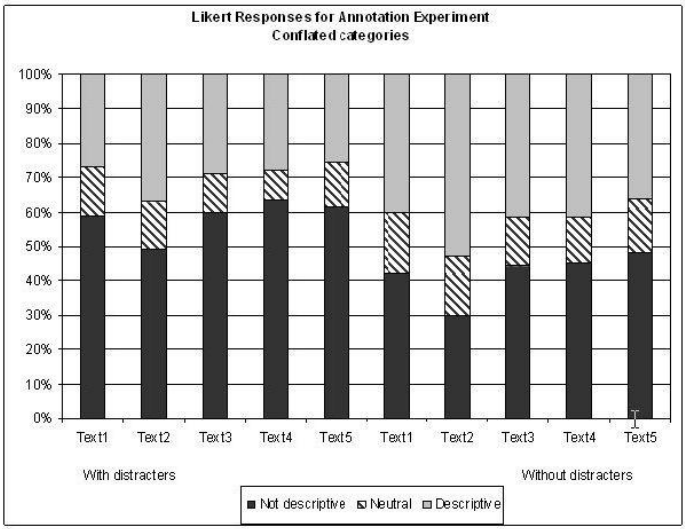


Figure 5: Conflated Likert Responses for Annotation Experiment

2 were conflated, as were categories 4 and 5). Each table gives kappa inter-rater reliability figures for three sets of data: all 15 terms (descriptors plus distracters), for ten descriptors, and finally for just the five distracters.

Table 3: Inter-rater agreement between Humans and CONSPECT

	free marginal kappa	without distracters	distracters only
Text 1	0.4	0.2	0.7
Text 2	0.4	0.3	0.5
Text 3	0.4	0.3	0.5
Text 4	0.4	0.3	0.8
Text 5	0.3	0.2	0.5
Average	0.4	0.3	0.6

Table 3 shows the highest agreement occurs when only the distracters are considered and the lowest agreement when the distracters are removed. Table 4 shows a similar pattern when conflated categories are examined. In each case (i.e. conflated and non-conflated categories) the reliability figure is lower than the accepted threshold of 0.7 (Randolph 2005) except when just the distracters were examined.
Finally, the agreement between humans and CONSPECT was evaluated. More specifically, the percentage of judgements where the humans gave a lower Likert rating for a

Table 4: Inter-rater agreement with cat. 1 and 2 and 4 and 5 conflated

	free marginal kappa	without distracters	distracters only
Text 1	0.5	0.4	0.8
Text 2	0.5	0.4	0.7
Text 3	0.6	0.4	0.8
Text 4	0.6	0.4	1.0
Text 5	0.5	0.4	0.7
Average	0.5	0.4	0.8

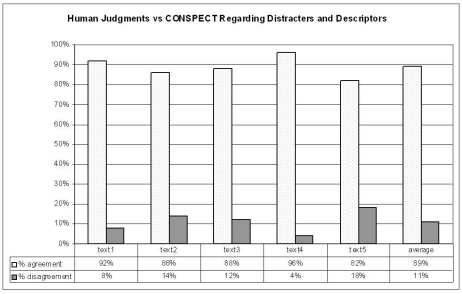


Figure 6: Comparing Human and CONSPECT Judgments

distracter compared to each descriptor was calculated. Figure 6 shows that the average agreement was 89%. This finding, along with that shown In Table 3 and Table 4, leads to the conclusion that CONSPECT is better at identifying whether a concept *is not* descriptive than it is at deciding whether a concept *is* descriptive.

4 Future Work

Various further investigations are planned to improve the results, e.g., finding a better clustering algorithm. And as always when using LSA, decisions need to be made regarding thresholds, corpora, dimensions, and types of pre-processing (Haley, Thomas et al. 2007). In terms of pre-processing, using lemmatisation instead of stemming could be investigated. Another area ripe for research is using suffix array analysis for phrases; phrases are ignored completely in this implementation. In addition to changes to the LSA algorithm, other techniques can be investigated, such as explicit semantic analysis (Gabrilovich and Markovitch 2007) and higher order collocation analysis (Keibel and Belica 2007).

5 Conclusion

The overall conclusion, based on the results of these several analyses, is that CONSPECT shows enough agreement with humans that it is a good start for a system to monitor conceptual development. The previous section describes some of the possible improvements to be researched. In addition, the verification experiments will be repeated with Dutch Psychology students. This will provide very interesting data about how well CONSPECT works with a different language in a different knowledge domain.

6 Acknowledgments

CONSPECT was developed as a part of the Language Technologies for Life Long Learning (LTfLL) project (see http://ltfll-project.org/). The LTfLL project is funded by the European Union under the ICT programme of the 7th Framework Programme (Contract number: 212578). We would like to thank the University of Manchester, specifically Alisdair Smithies, for help and support in this investigation.

References

Brandes, U. and Erlebach, T. (2005). *Network Analysis: Methodological Foundations*. Springer.

CLARIANA, R. B. and WALLACE, P. (2007). A computer-based approach for deriving and measuring individual and team knowledge structure from essay questions.

Evans, V. and Green, M. (2006). *Cognitive Linguistics: An Introduction*. Lawrence Erlbaum Associates.

Fruchterman, T. M. J. and Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software: Practice and Experience*, 21(11):1129–1164.

Gabrilovich, E. and Markovitch, S. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, Hyderabad, India.

Goldsmith, T. E., Johnson, P. J., and Acton, W. H. (1991). Assessing structural knowledge. *Journal of Educational Psychology*, 83(1):88–96.

Haley, D. (2008). Applying latent semantic analysis to computer assisted assessment in the computer science domain: a framework, a tool, and an evaluation.

Haley, D., Thomas, P., Roeck, A. D., and Petre, M. (2007). Measuring improvement in latent semantic analysis-based marking systems: using a computer to mark questions about html. In *ACM 9th International Australasian Computing Education Conference*. ACE '07 Proceedings of the ninth Australasian conference on Computing education - Volume 66 ISBN:1-920-68246-5.

Keibel, H. and Belica, C. (2007). Ccdb: A corpus-linguistic research and development workbench.

Landauer (2007). *Handbook of Latent Semantic Analysis*. Lawrence Erlbaum Associates.

Randolph, J. J., Thanks, A., Bednarik, R., and Myller, N. (2005). Free-marginal multirater kappa (multirater kappa free): An alternative to fleiss ' fixed- marginal multirater kappa.

Rugg, G. and McGeorge, P. (1997). The sorting techniques: a tutorial paper on card sorts, picture sorts and item sorts. *Expert Systems*, 14:80–93.

Saeed, J. (2003). *Semantics*. Introducing linguistics. Blackwell Pub.

Salton, G., Wong, A., and Yang, C. S. (1974). A vector space model for automatic indexing. Technical report, Cornell University, Ithaca, NY, USA.

Schvaneveldt, R. W., Durso, F. T., and Dearholt, D. W. (1989). Network structures in proximity data. volume 24 of *Psychology of Learning and Motivation*, pages 249 – 284. Academic Press.

Upchurch, L., Rugg, G., and Kitchenham, B. (2001). Using card sorts to elicit web page quality atributes. *IEEE Software*, pages 84–89.

Wild, F. (2010). Learning an environment: Supporting competence development in personal learning environments with mash-ups and natural language processing.

Wild, F., Haley, D., and Buelow, K. (2010). CONSPECT: monitoring conceptual development. In *Proceedings of the 9th International Conference on Web-based Learning (ICWL 2010)*.

Sabrina Wilske, Magdalena Wolska

Meaning versus Form in Computer-assisted Task-based Language Learning: A Case Study on the German Dative

We report on a study which investigated the effects of three types of feedback realized in instructional dialogues with a computer-based language learning system for German. The interaction was framed within a directions giving task and the linguistic form in focus was the dative case in prepositional phrases. The feedback types differed with respect to the focus they put on form versus meaning and the explicitness of feedback in response to learner errors. The results of the study suggest that a stronger focus on form is related to greater accuracy gains in using the form. The integration of incidental focus on form within a primarily meaning-based task increases accuracy as well, however to a lesser extent.

1 Motivation and research questions

One of the research objectives actively investigated in the second language acquisition (SLA) community is to determine what types of instruction are most effective for foreign language learning. Generally speaking, language instruction can give priority to formal aspects of the language or to meaning and content. Long (1991) proposed a distinction between three types of instruction in terms of emphasis on form versus meaning: Instruction may require the learner to focus on meaning, on form, or both at the same time. While Focus on Meaning (FonM) does not draw learners' attention to linguistic forms at all, Focus on Forms instruction (FonFs) focuses on forms in isolation, providing no or only limited meaningful context. Focus on Form instruction (FonF) tries to integrate meaning and form by drawing learners' attention to linguistic forms as they arise within primarily meaning-oriented interaction.

Focus on Form is often realized within communicative interactions which provide opportunities for the learner to produce comprehensible output as well as to modify their output in response to feedback, thereby stimulating learning (Long, 1981; Krashen, 1985; Swain, 1985). The communicative approach advocates the use of goal-oriented communicative activities, *tasks*, in foreign language learning (Long, 1991; Ellis, 2003). Communicative goals of tasks should be framed in real world situations which elicit the use of the developing language from the learners. Important definitional properties of tasks are (1) primary focus on meaning, (2) clearly defined communicative outcome, and (3) free use of linguistic forms which the learner chooses. The third point gives rise to a potential problem when, as part of a pedagogical strategy, a specific grammatical structure of a language is targeted: Because learners are free to use any forms they want, one cannot guarantee that they will use the forms of interest. Therefore *focused tasks* have been proposed as an attempt to integrate forms and meaning. Focused tasks are designed in such way that learners are likely to use a specific target structure thereby improving its mastery.

One of the factors contributing to the effectiveness of communicative interaction is the type of feedback learners get in response to non-target-like contributions: (1) explicit vs. implicit feedback, and (2) prompting for a correction or not. Instruction is considered explicit if it contains an explanation of the language phenomenon in question or asks learners to attend to particular forms in the target language. It is considered implicit otherwise (Norris and Ortega, 2001). Corrective reformulations of learner's utterances or their parts, so called *recasts*, provide implicit feedback without prompting for correction and thus do not disturb the task-level conversation. By contrast, *metalinguistic feedback* (comments or questions related to the error which do not explicitly provide the correct form) is explicit and thus temporarily shifts attention from meaning to form. While both of these feedback types have been previously investigated in the classroom context (see, for instance, (Lyster and Ranta, 1997)) there has been little research into the efficacy of different feedback types in computer-based dialogic language instruction. A previous study by Norris and Ortega (2001) comparing the efficacy of different types of instruction (primarily non-computer based) suggests a slight advantage of explicit instruction over implicit by showing that the former results in higher test scores. The same study suggests that FonF and FonFs have equivalent effects. Ferreira (2006) found that when a computer interface is involved, feedback which prompted learners to correct their error yielded more learning gains than feedback which provided the correct target form.

In this paper we report on a study which compared the effects of three types of computer-based dialogue activities which differ in terms of the degree of focus on form vs. meaning, the degree of explicitness of feedback, and correction prompting strategies, on the acquisition of foreign language structures. The activities were performed using a type-written computer-based dialogue system. The interaction with the system was framed within a directions giving task and the linguistic form in focus is the German dative case in prepositional phrases. Our research questions were:

(1) Does computer-based task-oriented interactive instruction help learners of German improve accuracy on the use of the dative case in prepositional phrases?

(2) Is there a difference in the effect of free (FonF) vs. constrained (FonFs) type-written production on the acquisition of the dative in prepositional phrases?

(3) Is there a difference in the effect of implicit feedback (recasts) vs. explicit (metalinguistic) feedback on the acquisition of the dative in prepositional phrases?

The pedagogical goal was two-fold: Learners should improve their communicative skills in the scenario and their control of the target structure. In this paper we report the results on the latter.

In general, the idea of computer-based dialogues for foreign language learning is not new. Computer assisted language learning (CALL) has been an active research area for many years. With the progress in language technology the number of intelligent CALL systems, allowing learners to use natural dialogue, has been growing; see, for instance, (Holland et al., 1998; Harless et al., 1999; Seneff et al., 2004; Johnson et al., 2004). However, most systems are not built with the goal of transferring findings or testing hypotheses from the field of SLA; see (Petersen, 2010) for one exception.

Gender	Nominative	Dative PP	Translation
Masc	**der** Laden	hinter **dem** Laden	behind the shop
Fem	**die** Mensa	hinter **der** Mensa	behind the canteen
Neut	**das** Cafe	hinter **dem** Cafe	behind the cafe

Table 1: German dative in a prepositional phrase

Outline This paper is organized as follows: Section 2 describes the scenario, the target structure, and the types of instruction we evaluated. In Section 3 the implemented dialogue system is briefly outlined. Section 4 presents the design of an experiment we conducted. Section 5 summarizes the results of the study. In Section 6 we discuss the findings and conclude.

2 The Approach

In line with the focused tasks method, for the communicative instruction we selected a grammatical form and a task such that the form is natural to use within the task scenario and such that the scenario is meaningful and useful for the learner. We introduce the form and the task below.

2.1 The target forms and the tasks

The form: Dative case in prepositional phrases Among other uses, the dative case in German is required as an object of certain spatial prepositions. The dative case in German is marked morphologically on the gender-specific determiner of a noun phrase as well as on adjectives and in specific cases on the head noun. Table 1 shows the nominative and dative case forms (emphasized in bold) for the three German genders. Most locative prepositions used for describing static spatial relations require dative, among others, *vor* ('in front of'), *hinter* ('behind'), *neben* ('next to'), or *zwischen* ('between') The directional prepositions *zu* and *bis zu* ('to, towards') also require dative. These prepositions can be elicited in a task involving spatial descriptions.

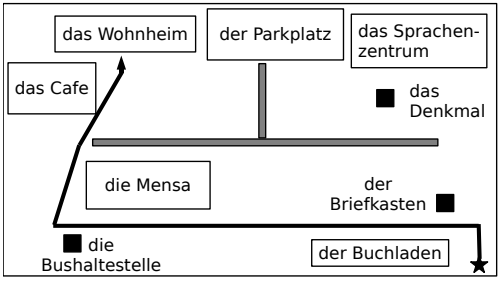


Figure 1: The map used in the "Directions giving" task

The task: Giving directions We designed the directions giving task in a way so that it most efficiently attempts to elicit the forms of interest. The learner is presented with a simplified map of a fictitious campus, with buildings, other landmarks and a route to describe. Figure 1 shows the actual material used in our study. The scenario described when presenting the task is that the learner was stopped on the campus and asked for directions. The instructions explicitly request that the map provided be used and that the indicated route be described. The task description does not include any hints as to using prepositional phrases or paying attention to the dative case. The landmarks we used are balanced as to their gender and the gender is provided on the map. The

Dialogue strategy

```
interpretation-found = false
get-user-input

if user-input-parsed
   interpretation-found = true
else
   if keyword-match-found
      interpretation-found = true

if interpretation-found == true
   if TF-realized
      generate-feedback-recast
         or generate-feedback-metaling
   else
      elicit-TF
else
   output
     `Sorry, I didn't understand.'
```

Feedback functions

```
generate-feedback-recast:
 if TF-incorrect
   recast-TF
 prompt-for-next-contribution

generate-feedback-metaling:
 if TF-incorrect
   if first-trial
      output
        `<DET> in <NP> is not correct.'
   else
      output
        `<DET> is not correct either.'
        `Use dative!
   prompt-for-correction
 else
   prompt-for-next-contribution
```

Figure 2: Dialogue strategy in the two free production activities as pseudo-code.

route includes two points of direction change each at two landmarks and the target is placed close to two other landmarks. This setup makes it likely that the dative will be used when referring to a turn at a landmark in order to make the directions more precise and it also gives an opportunity to ask clarification questions when the learner does not supply the target forms. The two landmarks at each point of direction change are either both feminine or masculine, while all the landmarks close to the target have neuter gender. The learner has thus an opportunity to use the dative with all the genders, but in case they do not use locative prepositions we can also explicitly elicit all of these forms.

2.2 Task-based instruction

We designed and implemented three variants of communicative instruction: All variants involved a type-written dialogue with the system we built in order to perform the "Giving directions" task. The system controlled the interaction by means of a state-based dialogue model. The three variants of the instruction differed in the extent of freedom of language production they offered and the realization of form-focused feedback. In two variants of FonF activities the learners were able to freely formulate their dialogue contributions, *free production*, while in the third variant, *constrained production*, implementing FonFs, the learners' production was limited to supplying the target form (filling a gap) . The feedback in the latter activity variant was explicitly stating whether the supplied form was correct or not. The two free production variants differed with respect to the feedback they provided in response to incorrect forms: One implicitly corrected the error while maintaining the focus on the task-level conversation whereas the other explicitly pointed the learner to the error and demanded a correction thereby briefly focusing on the form. We elaborate on the properties of the respective system variants below.

Free language production–FonF In the free-production FonF activities the learners were able to type their utterances freely without any restrictions on the language used. The system implemented two input interpretation strategies: one based on a grammar with mal-rules and the other a fall-back strategy, based on fuzzy keyword matching (see Section 3). The system classified the learner's input into one of three categories ("TF" stands for "target form"): TF-realized-correct, TF-realized-incorrect, TF-not-realized. The high-level dialogue and feedback strategy of the system is summarized as pseudo-code in Figure 2.[1] If the learner's input was classified as not realizing the target form, the system tried to elicit it once by asking a clarification request, as exemplified in (1):[2]

(1)	**L:**	und dann nach links	*and then left*
	S:	[wo soll ich links?]$_{ELICIT}$	*where do I turn left?*

In case of learner errors in the target form (the TF-realized-incorrect category) the recast system provided implicit feedback by reformulating, *recasting*, the learner's utterance or parts thereof. Recasts were realized in a way so as to give them an appearance of implicit confirmation type of grounding moves, as in (2).[3]

(2)	**L:**	Gehen Sie hinter **das** Cafe nach links.	*Turn left, past the coffee-shop*
	S:	Okay, [hinter **dem** Cafe nach links,]$_{RECAST}$	*Okay, left past the coffee-shop*
		[und dann?]$_{PROMPT}$	*and then?*

The metalinguistic feedback system would explicitly state that there is an error, point to the location of the error and elicit a correction by the learner, as shown in (3). In case the learner should not succeed in correcting the error, the system would give a further hint, as in (4); cf. Figure 2.

(3)	**L:**	Gehen Sie hinter **das** Cafe nach links.	*Turn left past the coffee-shop*
	S:	['das' in 'das Cafe' ist nicht richtig.]$_{METALING}$	*'das' in 'das Cafe' is not correct.*
		[Bitte noch einmal!]$_{PROMPT}$	*Please try again!*
(4)	**L:**	hinter **den** Cafe nach links.	*left past the coffee-shop*
	S:	['den' in 'den Cafe' ist auch nicht richtig.]$_{METALING}$	*'den' in 'den Cafe' is not correct either.*
		[Nimm Dativ!]$_{PROMPT}$	*Use the dative!*

The system did not attempt to diagnose nor correct any other incorrect structures except those in focus. We anticipated that some learners might give a complete route description in one turn at the start of the dialogue. In order to ensure longer engagement, the system prolonged the interaction

[1] We omitted some system turns signaling non-understanding due to unknown words in order to simplify the presentation.
[2] **S** and **L** mark system and learner turns respectively.
[3] The bold emphasis did not appear in system output and is used here only to indicate the incorrect form and its correction via recast.

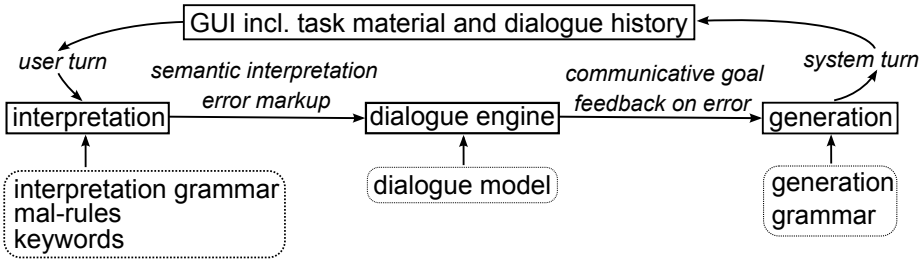


Figure 3: The system architecture; rectangles: modules, bottom part: resources, arrows: information flow

by asking the learner to slow down, confirming only the first part of the description, and prompting for continuation.

Constrained production–FonFs In the constrained production system, which implemented the strongly form-focused approach, the learner's production was restricted to supplying the target form by filling a gap in a pre-scripted dialogue turn as in the example below:

(5) **S:** Wie komme ich zur Mensa? *How do I get to the cafeteria?*
L: Gehen Sie hinter ☐ Cafe nach links. *Turn left past the coffee-shop*

The learner was allowed three attempts to produce the correct form. In case an invalid form was supplied, the system signaled it with a message 'That was wrong!' and subtracted one point from a learner's "score" on the activity; correct forms increased the score by one. The feedback and the score were displayed in a designated feedback area. After the third unsuccessful attempt the correct utterance was appended to the dialogue and the system generated its next turn.

This system was built on the same architecture as the free production systems. However, due to the constraint on the language production, it used a simpler method to map the input to the expected answer; case-insensitive exact string matching. The dialogue model was also simplified because the elicitation subdialogues and more elaborate feedback mechanisms were not employed.

3 The system

All three dialogue activities have been implemented on the same system architecture. In the description in this section, we concentrate on the components required for the free production activities; the constrained production activity is its a simplified variant.

The system maintains a dialogue with the learner by following the dialogue strategy outlined in Section 2.2 (see Figure 2). This involves interpreting the learner's input, responding to the learner by selecting a communicative goal according to the dialogue model and the pedagogical strategy, and realizing the goal as a surface string. Specifically for the learning context, the system has to recognize errors in the learner input (identify contributions in the TF-realized-incorrect category) and generate feedback on them.

```
<dir-change> = Gehen Sie <pp> nach (<left> | <right>)
<pp>         = <pp-DATIVE> {dat} | <pp-NODAT> {non-dat}
<pp-DATIVE>  = <prep> <np-dat>
<pp-NODAT>   = <prep> (<np-nom> | <np-gen> | <np-acc>)
```

Figure 4: A simplified fragment of the interpretation grammar including a mal-rule; `{non-dat}` is the semantic tag indicating that a dative PP was not used where it was expected.

Figure 3 shows the system's architecture: the modules, the resources they employ and the units of information that are passed between them.

The dialogue model and engine The dialogue model represents the sequences of possible turn transitions: alternating turns produced by the user and the system. It is implemented as a state machine using State Chart XML (SCXML) as an underlying representation. We use the Java implementation of Apache SCXML.[4] The Apache framework also provides a dialogue execution engine which receives input interpretations and triggers system responses according to the model.

Interpretation of learner's input In general, interpreting the user input involves mapping a surface string of an utterance to a meaning representation. As typical in small-scale dialogue systems, we implemented the system's language model (the set of linguistic expressions it covers) as a context free grammar with semantic tags. For parsing, we use the Java Speech API implementation of the CMU parser which is part of the Sphinx system.[5] The semantic tags encode two types of information: first, the symbolic meaning of utterances, and second, information on violations of grammatical constraints. Two error handling strategies are implemented in the system:

Fuzzy matching for unknown words In order to ensure robustness with respect to typos and spelling errors the system first identifies unknown words in the input and tries to map them to known words by calculating the Levenshtein distance between the unknown word and known words. For replacement with in-vocabulary candidates we consider those words which have a Levenshtein distance within a certain range to a known word, normalized by word length.

Grammatical error handling Since the system interacts with learners, i.e. non-native speakers of German, their input is likely to contain other errors apart from misspellings, in particular errors in the target structure. An essential requirement of the system is to recognize those errors and give feedback on them. One strategy to deal with errors is to explicitly integrate anticipated errors into the grammar in the form of so called mal-rules, i.e. grammar productions which are outside of the standard rules of the given language. Erroneous utterances are parsed using mal-rules and the parse result contains information about the error. Figure 4 presents a fragment of the interpretation grammar, including mal-rules. The rule `<dir-change>` covers the utterance given in (2). If the prepositional phrase `<pp>` is not in the dative case, the semantic tag `non-dat` is returned, indicating that the dative case was required, but was not found. We encoded a set of mal-rules based on informal prior pre-testing of the system with beginner learners.

[4] http://commons.apache.org/scxml
[5] http://cmusphinx.sourceforge.net

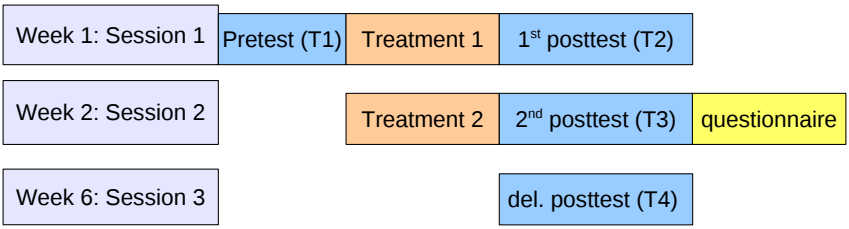


Figure 5: Experiment timeline

The drawback of this approach is that it is hard to anticipate all possible errors that might occur. Therefore, our system also implements a fall-back strategy based on keyword spotting: If no parse is found for an utterance, we create a semantic interpretation based on content words, using a keyword lexicon.

Generation of system responses The system output realization is performed using a template-based approach. The output is produced by generating a dialogue move selected according to the dialogue model using a context free generation grammar. The grammar associates atomic symbols representing communicative goals with sets of possible realizations. The generation templates contain slots encoding references to landmarks or directions for confirmation moves, or grammatical information for error feedback. Slots in the templates are filled using feature-value pairs passed as arguments to the templates along with the communicative goals to be realized.

4 The Experiment

In order to answer our research questions we conducted an in-classroom quasi-experimental study with the systems we built, in a pretest multiple-posttest design. The setup of the experiment is presented in the following sections.

4.1 Design and procedure

The study used a quasi-experimental design involving 60 students from six German language classes (ranging from A2 to B1+ CEF level (Trim et al., 2001)), taught by different teachers. The courses met twice a week for 90 minute sessions. The experiment took place six weeks (approximately 15 instruction hours) into the course. In each class, participants were randomly assigned to one of the three conditions: free production-recast, free production-metalinguistic feedback, and constrained production. Figure 5 illustrates the timeline and setup of the procedure. The experimental groups participated in two sessions of the computer-based communicative instruction with one week's break between the sessions. Each session consisted of at least two repetitions of the activity in different configurations of the task material.

At the first session, all groups completed a pretest (T1, see below) and worked with the system (the treatment) followed by a posttest (T2). Another posttest (T3) followed the second session

of the treatment, and another posttest (T4) after a five week break. After the second session the subjects completed a short questionnaire on biographical information and feedback on the system.

4.2 Tests

We used two types of tests in the study: an *untimed sentence construction* test, targeting explicit knowledge, and a *timed grammaticality judgment* test, targeting implicit knowledge. Explicit knowledge is knowledge accessible through controlled processing, while implicit knowledge is accessible through automatic processing, i.e. learners' intuitive awareness of the linguistic norms (Ellis et al. (2006)).[6]

Timed grammaticality judgment Following Ellis (2006) we designed a timed grammaticality judgment test to measure implicit knowledge. The test items included different combinations of six different spatial prepositions (*bei* ('at'), *hinter* ('behind'), *neben* ('next to'), *vor*, ('in front of'), *zu* ('to'), *auf* ('on')) and nouns of the three genders, equally balanced. The underlying problem with testing the dative case is that learners need to know the gender of the noun in order to make a judgment about the correctness of a prepositional phrase. Because we did not want to test the learners' knowledge of genders, but their knowledge of the datives, we chose common feminine and masculine nouns whose grammatical gender matches the semantic gender, e.g. mother, man, son, etc. For neuter nouns we chose words that are usually taught at the beginner's level, e.g. child, horse. However, due to logistic constraints, we could not explicitly test whether the gender of the nouns included in the test items were indeed known.

The test included 9 grammatical, 8 ungrammatical test items[7] and 7 grammatical and 7 ungrammatical distractor items. The time-limit was set to 10 seconds per item. (This is roughly twice the maximum time a native speaker used).[8] Each correctly judged item was scored at 1 point, each incorrectly judged item was scored at 0.

Sentence construction For the explicit knowledge test, participants were asked to complete sentences given the beginning of a sentence and a set of unordered uninflected phrases or words. Full noun phrases were given along with gender information, as in the example below:

Item: Das Pferd (stehen, die Kuh, vor)
Solution: Das Pferd steht vor der Kuh. *The horse stands in front of the cow.*

The test consisted of 8 test items containing 6 prepositions (*bei* ('at'), *hinter* ('behind'), *neben* ('next to'), *vor*, ('in front of'), *zu* ('to'), *zwischen*, ('between')) with a gender-balanced set of nouns, and 4 distractor items.[9] There was no time-limit on the test items. The item was scored 1

[6] The tests were prepared and administered using Webexp Experimental Software. `http://www.hcrc.ed.ac.uk/web_exp/`

[7] One of the original 9 ungrammatical items was disregarded in the evaluation because of a spelling error we overlooked.

[8] Ellis timed his test at 20% above the average time native speakers required (Ellis, 2006). Han and Ellis (1998) used 3.5 seconds as the time constraint based on pretesting the items, while Bialystok (1979) used an even shorter time limit. Based on our pretest, already the threshold of 3.5 seconds would have excluded a couple of slow native speakers. Since we are not aware of research which explicitly addresses the issue of the time limit on the timed judgment tasks, we opted for a more generous time-limit.

[9] Note that the used prepositions differ slightly between the two tests types for practical reasons: For instance, although 'between' is a relevant preposition, we did not use it in the grammaticality judgment test, because it requires two noun

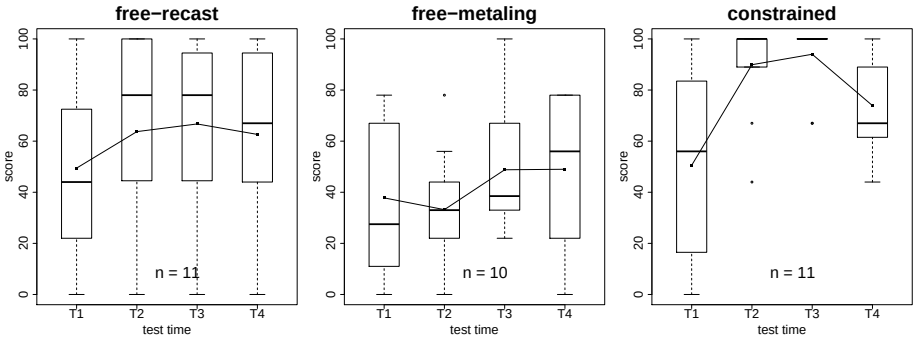


Figure 6: Results for sentence construction test.

point if the prepositional phrase was built correctly. The item with the preposition 'between' was scored at 1 point for each correct noun phrase. All other form errors were neglected.

We created four versions of each of the tests to be administered at the four times of assessment (T1, T2, T3, T4). The versions differed in the combinations of prepositions and noun phrases, but were otherwise comparable with regard to lexical items used. The assignment of a test version to a test time was randomly varied for each participant in order to compensate for any unintended differences between test versions. Within each test items were presented in random order.

4.3 Analysis

With the experiment spanning over six weeks, subject drop-out was inevitable. Due to a high drop-out rate (around 50%) and incidental data loss we have data for only 32 subjects for the sentence construction test, and 30 subjects for the grammaticality judgment test, around 10 for each experiment condition. We performed non-parametric analyses because of the small sample size and because parametric assumptions were not met: According to Shapiro-Wilk and Levene tests, both the normality assumption and the assumption of homogeneity of variance were violated on at least some of the within-subject and/or between-subject variables on either tests.

In order to compare within subject differences we performed Friedman tests followed by pairwise post-hoc comparisons using Wilcoxon signed rank test on those groups for which the Friedman test was statistically significant. For between-group comparisons we used the Kruskal-Wallis one-way analysis of variance test, followed by pairwise post-hoc comparisons using the Mann-Whitney U test. The significance level was set at 0.05. We mark differences which were significant at $\alpha = 0.10$ to indicate interesting tendencies.

		T1		T2		T3		T4	
Group/test	n	m	sd	m	sd	m	sd	m	sd
Free-recast									
SC total	11	49	32	64	39	67	35	63	38
TGJT total	11	63	23	78	22	79	18	76	17
TGJT gram.		76	21	83	18	84	15	88	13
TGJT ungr.		49	29	73	33	73	27	62	31
Free-metaling									
SC total	10	38	31	33	22	49	26	49	32
TGJT total	9	52	18	62	22	63	20	64	15
TGJT gram.		63	22	77	22	78	16	83	14
TGJT ungr.		40	29	46	33	46	34	43	27
Constrained									
SC total	11	51	36	90	18	94	13	74	19
TGJT total	10	68	19	90	16	87	13	81	18
TGJT gram.		78	21	92	10	89	12	89	14
TGJT ungr.		58	27	88	24	85	18	71	26

Table 2: Test results: means (m) and standard deviations (sd) for percentage scores

5 Results

The analyses below are based on the data set for the 30 (or 32) subjects with test results for all four assessment times. Table 2 shows the mean percentage scores and standard deviations for each experimental group on both tests: sentence construction (SC) and timed grammaticality judgment test (TGJT). For the latter test, the table also shows the scores for grammatical and ungrammatical items separately.

5.1 Sentence construction test

Figure 6 shows box plots and means for the sentence construction test. The first point to note is that there was no significant between group difference on the pretest on both tests. This means that before the treatment the groups were at the same level. The free-recast group and the constrained production group both increased from pretest (T1) to the first posttest (T2) and slightly further increased between from the first posttest (T2) to the second posttest (T3). The free-metalinguistic group slightly deteriorated between T1 and T2, but improved between T2 and T3. All groups deteriorated at the delayed posttest (T4). Within-subject analysis of variance showed that there were significant differences in the scores across the three time periods in the constrained production

phrases that have to be judged at the same time, which makes it impossible to determine based on which the judgment was made.

group, but not in the other conditions. Post-hoc analysis showed that the constrained production group was significantly more accurate on the two posttests (T2 and T3) than on the pretest (T1).

Between-group comparisons showed significant difference at T2 and T3. Post-hoc analyses showed that at T2, the free-recast and the constrained production group were both significantly more accurate than the free-metalinguistic group.

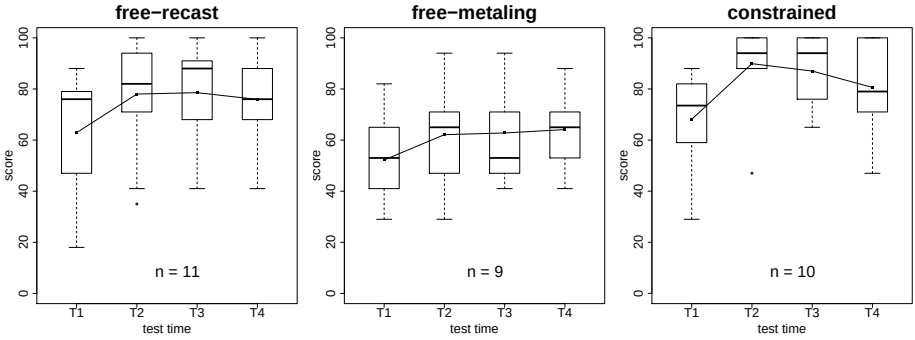


Figure 7: Results for timed grammaticality judgment test.

5.2 Timed grammaticality judgment test

Figure 7 illustrates the results for the timed grammaticality judgment test. As with the sentence construction test, the groups showed no significant difference at T1. All groups increased between T1 and T2. Between T2 and T3, only the constrained production group slightly deteriorated. The other groups showed no difference. All groups slightly deteriorated between T3 and T4. Within-subject analysis of variance showed that there were significant differences in the scores across the three time periods in the constrained production condition, with post-hoc analysis showing that the accuracy was significantly higher on T2, T3 and T4 than on T1. The free-recast group showed a difference across all test times that was significant at $\alpha = 0.10$, and further analysis showed that this group yielded a significantly higher score at T2 and T3 than at T1, ($p < 0.05$). The free-metalinguistic group showed no significant difference across test times. If we consider only the ungrammatical items of the grammaticality judgment test, these differences are maintained. However, for the grammatical items, the free-metalinguistic group shows a marginally significant difference ($p = 0.06$), with post-hoc analysis revealing that their score at T3 is higher than on T1 (marginally significant, $p = 0.06$) and their score at T4 is higher than on T1 ($p < 0.05$). The other two groups show no significant within-subject differences on the grammatical items.

Between group comparisons showed that there are marginally significant differences between groups at T2 and T3. Post-hoc analysis revealed that in both instances, the free-recast and constrained production group scored higher than the free-metalinguistic group.

groups	free-recast			free-metaling			constrained		
tests	●●○○	●●●○	●●●●	●●○○	●●●○	●●●●	●●○○	●●●○	●●●●
n: SC/TGJT	21/21	18/18	11/11	25/25	20/20	10/9	20/21	14/15	11/10
SC-total	(1-2)	-	-	-	1-3,(2-3)	-	1-2	1-2,1-3	1-2,1-3
TGJT-total	1-2	-	1-2,1-3,1-4	1-2	1-2,(1-3)	-	1-2	1-2,1-3	1-2,1-3
TGJT-gram.	-	-	-	-	-	(1-3),1-4	1-2	1-2	-
TGJT-ungr.	1-2	1-2	1-2,1-3	1-2	1-2	-	1-2	1-2,1-3	1-2,1-3,(1-4)

Table 3: Significant changes between test times for each group and each subset of tests taken.

5.3 Additional results for subsets of assessment times

Given the small number of subjects leading to low power of the tests, we also conducted analyses with data comprising only the first two or three tests respectively, for which the data of 67 (or 53, respectively) subjects is available. Table 3 shows for which subsets of assessment times there were significant within- and between-subject differences. For each of the three conditions the table shows three columns indicating the data of all subjects taking part in the first two (●●○○), the first three (●●●○) , or all four tests (●●●●) respectively. The values in the table cells indicate between which of the respective tests there was a significant difference. Brackets indicate that the difference is only significant at $\alpha = 0.10$. For instance, for the free-metalinguistic group on the timed grammaticality judgment test (TGJT-total), for all subjects taking part in all tests (●●●●) , there was a significant difference between T1 and T4 (at $\alpha = 0.05$) and a marginally significant difference between T1 and T3 (at $\alpha = 0.10$) .

While we found similar within-subject differences for the constrained production group, most interestingly, more differences became evident for the free production groups. More specifically the free-metalinguistic group showed significant improvement on the sentence construction test between T1 and T3 (and between T2 and T3 at $\alpha = 0.10$). This group also showed an increase in accuracy on ungrammatical items between T1 and T2. The free-recast group showed an increase in performance in the sentence construction test between T1 and T2 (at $\alpha = 0.10$) when only considering data for these two assessment times.

6 Conclusion

We presented a study which investigated the efficacy of different computer-based form-focused task-oriented activities on the acquisition of the German dative in a certain type of prepositional phrases. Noting that the number of subjects whose data we were able to analyze statistically was rather small, the implications of this study should be taken cautiously. Based on the analyses, certain tendencies can be however observed.

First, not surprisingly, most of of the effect is found between the pretest and the first posttest, that is, there is an immediate effect of the intervention. Second, also not surprisingly, the explicit Focus on Forms instruction (constrained production) appears to achieve more of the effect.[10] It

[10] Considering the drill-like character of the constrained production dialogues, it would be of course interesting to contrast it with a simple traditional decontextualized drill in order to see whether there is any added value to the embedding in the dialogue interaction.

appears that the learning in the free production conditions is slower (stepwise increase in the mean scores in the free production groups vs. a jump of the scores in the constrained condition). We cannot draw a clear conclusion to our third research question, whether there is a difference between the different feedback types. In general, the recast group achieves more significant gains in accuracy when taking into account the data for all four test times. However, if we do not consider the delayed posttest, the metalinguistic group seems to achieve the same effect on a larger data set.

It is interesting that the free-recast group achieves more of the significant results on the implicit knowledge test than on explicit knowledge. This might be due to, on the one hand, indirect nature of the feedback and a weaker form-focusing mechanism than in the other condition, and on the other hand, due to stronger engagement in the activity and, possibly, better noticing of feedback (recasts) as a result.

While the presented analysis focused on accuracy in the usage of the target structure as the only measure of language development, we also tested the effect of the task activities on spoken language fluency on an analogous task. In the beginning of each session participants were asked to work in pairs and describe a route on a map. The ensuing conversations were recorded. For the subset of the data – 13 participants of the constrained production and the free-recast condition – we analyzed the transcripts of the speech samples with regard to durational measures associated with fluency. In addition, we also asked German teachers to rate and rank those samples with respect to the perceived fluency. However, the results were not clear cut. When correlating the ranking of raters with the test times, a slightly higher positive correlation was evident for the free-recast group than for the constrained production group. On some durational measures, the free-recast groups improved significantly while the constrained production group showed no difference. For other measures it was the other way round. However, given the small number of subjects, again these results have to be taken cautiously.

As part of future work, we are planning to analyze the accuracy in the use of the target structure within the oral test as well as in the system interaction dialogues. We are presently annotating the interaction data (the system logs) along two dimensions: the grammatical aspects of the learner language and the structure of the interaction, in order to be able to investigate interaction-based correlates of the results we presented in this paper.

Acknowledgments

We would like to thank Dr. Kristin Stezano Cotelo and Meike van Hoorn, language instructors at the German courses of the International Office at the Saarland University, as well as their students who participated in the experiments, for the help in conducting this study.

Sabrina Wilske's work was funded by the IRTG PhD program "Language Technology and Cognitive Systems". Magdalena Wolska's position at Saarland University is partially funded through the INTERREG IV A programme[11] project ALLEGRO (Project No.: 67 SMLW 1 1 137).

[11] http://www.interreg-4agr.eu/

References

Bialystok, E. (1979). Explicit and implicit judgements of l2 grammaticality. *Language Learning*, 29:81 – 103.

Ellis, R. (2003). *Task-based Language Learning and Teaching*. Oxford University Press.

Ellis, R. (2006). Modelling learning difficulty and second language proficiency: The differential contributions of implicit and explicit knowledge. *Applied Linguistics*, 27(3):431–463.

Ellis, R., Loewen, S., and Erlam, R. (2006). Implicit and explicit corrective feedback and the acquisition of l2 grammar. *Studies in Second Language Acquisition*, 28:339–368.

Ferreira, A. (2006). An experimental study of effective feedback strategies for intelligent tutorial systems for foreign language. In *IBERAMIA-SBIA*, pages 27–36.

Han, Y. and Ellis, R. (1998). Implicit knowledge, explicit knowledge and general language proficiency. *Language Teaching Research*, 2:1–23.

Harless, W., Zier, M., and Duncan, R. (1999). Virtual dialogues with native speakers: The evaluation of an interactive multimedia method. *Calico Journal*, 16(3):313–37.

Holland, V. M., Kaplan, J. D., and Sabol, M. A. (1998). Preliminary tests of language learning in a speech-interactive graphics microworld. *Calico Journal*, 16(3):339–359.

Johnson, W., Marsella, S., and Vilhjaálmsson, H. (2004). The DARWARS tactical language training system. In *Proceedings of the Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC)*.

Krashen, S. D. (1985). *Input Hypothesis: Issues and Implications*. Longman.

Long, M. H. (1981). Input, interaction and second language acquisition. In Winitz, H., editor, *Native language and foreign language acquisition*, volume 379, pages 259–78. Annals of the New York Academy of Sciences.

Long, M. H. (1991). Focus on form: A design feature in language teaching methodology. In Bot, K., Ginsberg, R. B., and Kramsch, C., editors, *Foreign language research in cross-cultural perspective*, pages 39–52. John Benjamins.

Lyster, R. and Ranta, L. (1997). Corrective feedback and learner uptake: Negotiation of form in communicative classrooms. *Studies in Second Language Acquisition*, 19:37–66.

Norris, J. M. and Ortega, L. (2001). Does Type of Instruction Make a Difference? Substantive Findings From a Meta-analytic Review. In Ellis, R., editor, *Form-Focused Instruction and Second Language Learning*, pages 157–213. Blackwell.

Petersen, K. (2010). *Implicit corrective feedback in computer-guided interaction: Does Mode Matter?* PhD thesis, Georgetown University.

Seneff, S., Wang, C., and Zhang, J. (2004). Spoken Conversational Interaction for Language Learning. In *Proceedings of INSTIL/CALL*, pages 151–154, Venice, Italy.

Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In Gass, S. and Madden, C., editors, *Input in second language acquisition*, pages 235–53. Newbury House.

Trim, J., North, B., and Coste, D. (2001). *Gemeinsamer europäischer Referenzrahmen für Sprachen: Lernen, lehren, beurteilen. Niveau A1, A2, B1, B2*. Langenscheidt.

Haruko Miyakoda, Kei-ichi Kaneko, Masatoshi Ishikawa

Effective learning material for mobile devices: Visual data vs. Aural data vs. Text data

According to previous studies, visual data help to enhance vocabulary learning in a foreign language (e.g. Yeh and Wang 2003). However, it is not easy to create visual data for all lexical items. Particularly, we encounter problems when dealing with abstract words.

In this paper, we report on two experiments that tested the learning effects of four types of materials for mobile devices that were based on the following contents: 1. translation only, 2. aural data only, 3. visual data only, 4. aural and visual data. In Experiment 1, the 59 subjects were instructed to memorize the lexical items that were stored in 4 different types mentioned above, and vocabulary tests were administered in order to measure the memory retention rate of the meaning of each word.

Contrary to previous findings, the material employing aural data only drew out the best results. In addition, the material employing just translation scored better or did as well as either the translation + visual data or the translation + visual + aural data. This implies that contrary to claims found in the literature, visual data may not be so significant a factor in vocabulary learning. However, the difference was not statistically significant. In order to test the validity of this result, we conducted Experiment 2, following similar procedures. 40 subjects participated in the experiment. This time, again, visual data did not seem to provide much aid in facilitating vocabulary attainment, since the mean test scores were similar between translation only and translation + visual material.

The results from these two experiments suggest that contrary to claims in the literature, we may not need to rely too much on visual data in vocabulary attainment. Furthermore, material based on text data only may also prove to be an effective means of learning vocabulary in a foreign language.

1 Introduction

There has been a revival of interest in vocabulary teaching in recent years. This is partly due to the development of new approaches to language teaching, such as the lexical approach (Thornbury 2002). In addition to this, there are research findings showing how lexical problems can cause serious communication breakdowns, more severe is the nature of these problems than has been pointed out in the literature (Allen 1983).

Unlike the learning of grammar, vocabulary is largely a question of accumulating individual lexical items into long-term memory (Thornsbury, 2002). This means that one of the successful ways of achieving vocabulary attainment is to spend time on repetitive memorization activities (Schmitt and McCarthy, 2005). In this sense, ubiquitous autonomous learning can be seen as an ideal method of learning vocabulary, allowing learners to increase the

time of exposure to the words to be learned, thus making good use of their time outside the classrooms.

With the advent of computers, new tools for studying vocabulary have been presented. Particularly, E-Learning based on mobile devices is getting more and more popular as a way of learning a foreign language (Amemiya et al., 2007). Employing mobile devices in vocabulary learning is an ideal way of studying because the mobility and portability of these devices provide the users with a ubiquitous environment, where they can study whenever and wherever they like. In addition to providing a ubiquitous environment, there is, of course, also the need to consider carefully the content of the learning material that is employed. This will be the topic to be taken up in the next section.

2 Learning material

Many papers dealing with learning material can be found in the literature, mostly supporting the effectiveness of visual data in facilitating the vocabulary learning process. For example, some studies investigating the difference between annotations by still images and those by movies, conclude that the learning effect by the annotations based on movies and texts are superior to those by still images and texts (cf. Al-Seyghayar 2001). On the other hand, other studies conclude that the annotations based on texts and still images are most effective (cf. Yeh and Wang 2003). Although these studies each have come up with different results as to whether movies or still images are effective, they all agree that visual data play an important role in vocabulary acquirement. However, as can easily be expected, although visual data may be effective, not all lexical items can be expressed visually. Furthermore, even if a word could be expressed using visual data, it does not necessarily mean that everyone will come up with the same visual image for the same lexical item.

In order to find out what role visual data play in vocabulary attainment, we conducted an experiment that compared the learning effects of different material, the details of which will be given in the following section. Before going into details of the experiment, we will briefly outline the vocabulary learning online system that we employed in the experiment.

3 Outline of the system

The online vocabulary system that we have developed consists mainly of three subsystems: 1. a system that supports or facilitates the creating process of the learning materials for mobile devices (Personal Super Imposer), 2. a system that supports its users in downloading the entities from the database and storing them for personal use (Personal Handy Instructor, 3. a system that allows users to share and evaluate the learning entities among themselves (SIGMA). Our emphasis in developing the system was to enhance the use of mobile devices in language learning and also to have the learners participate in creating the materials rather than just passively employ what has been prepared for them, because active involvement is expected to lead to effective learning.

3.1 Personal Super Imposer

As mentioned above, Personal Super Imposer (PSI) is a subsystem that supports its users in creating vocabulary learning material for mobile devices. When PSI is fed a five-second movie clip and the corresponding text data of the lexical item to be learned (namely the spelling of the word and its meaning), it automatically creates a multimedia learning entity. The process is outlined in Figure 1:

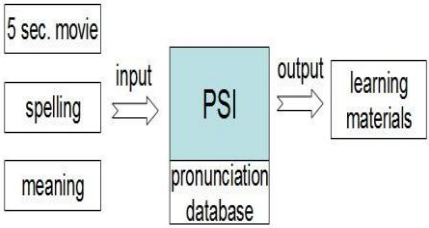


Figure 1: Personal Super Imposer

In Figure 2, a sample of the material made by PSI is given. The length of the movies for each entity is five seconds long. This length was determined based on the result of our pilot study that indicated that it was adequate in allowing most learners to process the visual/aural data repeatedly without feeling stress.

The pronunciation of each word is repeated twice in one learning entity. The spelling is displayed from the beginning of the entity but the corresponding meaning in the native language appears two seconds later. This time gap is necessary in order to give the learner a chance to concentrate first on the spelling before trying to memorize the meaning.

Figure 2: Sample of learning material (English-Japanese)

One of the advantages of creating material with PSI is that the same material can be reused or recycled so that it can be applied to virtually any language or dialect. For example, the item given in Figure 2 was originally created for Japanese learners of English. That is, the English word (foreign lexical item) appears on the first line and the corresponding meaning in Japanese appears on the second. If we change the typed-in information from Japanese to Chinese, the system automatically transforms the entity into an English-Chinese item (Figure 3):

Figure 3: Sample of learning material (English-Chinese)

3.2 Personal Handy Instructor

Personal Handy Instructor (PHI) is a vocabulary learning system for mobile devices such as iPods and mobile phones. PHI employs the five-second movie created by PSI mentioned above as its learning material. Figure 4 shows the process of learning vocabulary using PHI.

First, the learner selects the learning material that he/she wants to use from the learning-material list managed by PHI. The chosen material is copied into a folder called a 'vocabulary book.' Then, the users import the learning materials to their mobile devices such as iPods by dragging and dropping the folder onto the iTunes window. Finally, the users can download the learning material from iTunes to iPod. An outline of this process is summarized in Figure 4.

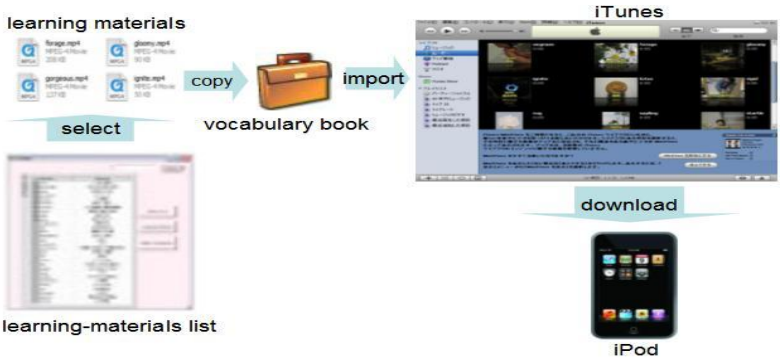


Figure 4: Personal Handy Instructor

3.3 SIGMA

The SIGMA system is a Web application that uses Apache, PHP and MySQL. It was designed to let learners register their own learning material and download the material created by other users.

In addition, it allows the users to evaluate the learning material and also give comments to each one. Figure 5 shows the main frame of the SIGMA system

If a user just wants to browse through the evaluation scores or comments of the learning materials, no login operations are required. However, if users want to evaluate the learning material or give comments on them, they need to become authorized users. Only authorized

users can register and manage their own material and give evaluation scores and comments back on all material after login operation.

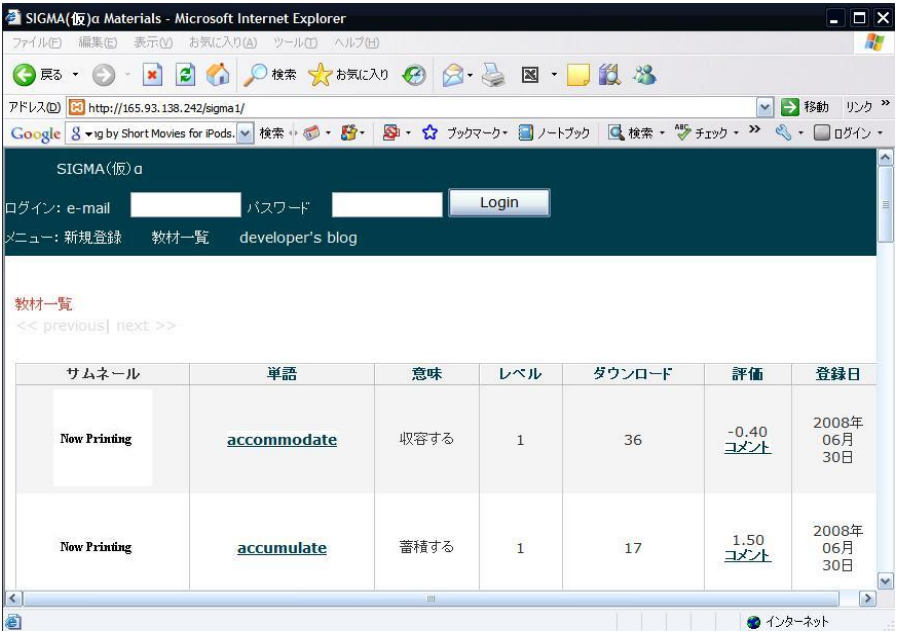


Figure 5: Main frame of the SIGMA

4 The Experiment No. 1

In order to see which factor is most effective in vocabulary learning, we conducted an experiment employing the online vocabulary system mentioned in the previous section.

4.1 The procedures

The main purpose of the experiment is to test the efficacy of visual data in vocabulary learning. Particularly, we wanted to compare the following four different types of material: 1. translation only (hereafter tr), 2. translation + aural data (hereafter aur), 3. translation + visual (hereafter vis), 4. translation + visual + aural (hereafter vis+aur). 59 undergraduate and graduate students attending a university in Tokyo, Japan participated in the experiment. The procedures prior to the experiment are as follows:

1. A vocabulary test was conducted by all participants in order to distinguish the lexical items that they were familiar with from the ones that were least familiar. The items that were least familiar among the participants were considered to be candidates for the experiment.

2. Based on the result of the vocabulary test in 1, we selected the following 15 items for use in the experiment: *ajar, beckon, bib, bicep, detour, disheveled, diverge, faucet, gargle, glimpse, hibernate, lament, perspire, pollen, stroll.*
3. The learning material for the 15 words above was created using the PSI system. For the (tr) and (aur) entities, no visual data is provided, and the screen would look something like Figure 6, where only the English word and the corresponding Japanese translation are provided as subtitles on a blank screen. The only difference between the (tr) and (aur) entities is that for the (aur) material, the pronunciation of the English word is additionally provided by sound data.

Figure 6: Example of iPod screen for (tr) and (aur) materials

4. For the (vis) and the (vis + aur) entities, in addition to the original English word and its translation in Japanese, visual data corresponding to the meaning of the word is provided, as the following example in Figure 7 indicates.

Figure 7: Example of iPod screen for (vis) and (vis+aur) materials

The procedures for the experiment itself are as follows:

1. The subjects were randomly divided into four learning groups (Group A-D), according to the type of material they were assigned to employ (i.e. Group A, (tr); Group B, (aur); Group C, (vis); and Group D, (vis+aur)).
2. The subjects were administered a test (Test 1) which included the 15 words mentioned above. In the test, the subjects were asked to write down the translation of the 15 lexical items in Japanese.
3. The subjects were given 5 minutes learning time. They were instructed to memorize the meaning of the 15 words using mobile devices.
4. Test 2 was conducted, which was based on English to Japanese translation tasks.
5. Test 3 was conducted a week later. The task involved was the same as Test 2.

4.2 Result and Discussion

The result of the experiment is summarized below in Table 1.

Test 1	A	B	C	D	Total
aver.	0.73	0.47	1.00	0.13	**0.58**
SD	1.87	1.13	2.32	0.52	**1.59**
Var.	3.50	1.27	5.38	0.27	**2.52**
Test 2					
aver.	21.20	22.47	20.00	20.20	**20.98**
SD	5.88	6.30	4.62	6.71	**5.88**
Var.	34.60	39.70	21.38	45.03	**34.57**
Test 3					
aver.	11.33	13.40	11.73	11.14	**11.96**
SD	7.45	6.63	7.94	5.60	**6.73**
Var.	55.52	43.97	63.02	31.36	**45.29**

Table 1: Result of Experiment 1

The full score for each test was 30 points (2 points x 15 lexical items). The participants received 2 points if they were able to answer the meaning of the lexical item correctly, 1 point if the meaning was partially right (e.g. “to be hot” for “perspire”), and 0 point if they were not able to come up with the correct answer.

When we focus on Test 1, we find that the results among the four groups vary slightly, with Group C having the highest average score of 1.00 out of 30.00, and Group D having the lowest score of 0.13. The average for all 59 subjects came out as 0.58. Since the 15 words employed in the experiment were all supposed to be least familiar to the subjects, the low average score for Test 1 was what we had expected. When we shift our attention to Test 2, which was administered to the participants right after the 5 minute learning process, we find that Group B, which employed the (aur) material, had the highest average score of 22.47 out of 30.00, followed by 21.20 for Group A (tr), 20.20 for Group D (vis+aur), and 20.00 for Group C (vis). For Test 3, which was administered to the participants one week after the learning process, here again, Group B had the highest average score of 13.40 out of 30.00, followed by Group C (11.73), Group A (11.33) and then Group D (11.14). In both Tests 2 and 3, the group that came out with the highest average score was Group B, which had used the translation and the sound data to memorize the meaning of the words.

In addition to these findings, we calculated the retention rate of the words' meanings between Tests 2 and 3. The retention rate was obtained by comparing each participant's test score for Test 2 with that of Test 3.

The overall results are summarized below:

Group A (tr) 52.70%
Group B (aur) 57.90%
Group C (vis) 52.53%
Group D (aur + vis) 57.25%

Here again, Group B scored the highest, followed by Group D, Group A, and then Group C.

In all cases, the material employing aural data only drew out the best results. Interestingly, in addition, the data employing translation only scored better or did as well as the visual data. This implies that contrary to claims found in the literature, visual data may not be playing as crucial a role as one might expect. As mentioned above, previous studies related to vocabulary acquirement generally emphasize the importance of employing visual data for effective learning. However, this is easier said than done. For one, it is very time-consuming to create visual data for use in the classroom. Finding the right visual data that corresponds to the word is a bit of a burden, and even with the aid of technology, it still requires much time and effort in the creation process. A further problem arises in the case of abstract words; visual images are hard to create for these words to begin with. You cannot easily find images for words like "lament", as you would for words like "lollypop" or "tiger." Therefore, even though material employing movies and visual images may seem useful for vocabulary learning, we must not forget that it has its limitations. It is hardly practical for advanced learners, who, especially must cope with abstract terminology most of the time. The result of our experiment indicates that there may not be the need to rely on visual data, and that employing either aural data only or translation only may be effective ways in vocabulary attainment. However, the results were not statistically significant. In order to verify whether visual data do play a role in vocabulary acquisition, we did a follow-up experiment, whose details will be given in the next section.

5 The Experiment No.2

40 university students participated in this experiment. They were randomly divided into four groups. Just as for Experiment 1, the four groups were divided according to the type of material employed in the learning process. That is, translation only (Group A), translation + aural (Group B), translation + visual (Group C), and translation + visual + aural (Group D). This time, however, in order to test the effectiveness of repeated learning, we conducted two sets of pre-test and post-test. In other words, after conducting the first set of pre-test and post-test on day one, a second round of the pre-test and post-test was conducted on day two, which was one week after day one.

The following 15 items employed in this experiment: *crumble, dilute, dormant, evalcuate, evaporate, glare, immerse, meditate, nudge, provoke, reconcile, sacrifice, seize, tempt, wither*.

The results of the mean scores for these two sets are given below:

pre-test 1	A	B	C	D	Total
	1.1	2.1	1.1	1.7	**1.5**
post-test 1					
	24.2	18.7	24.3	19.8	**21.7**
pre-test 2					
	13.7	8.6	13.8	9.9	**11.5**
post-test 2					
	29.3	26.8	29.7	28.4	**28.5**

Table 2: Mean Scores of Experiment 2

Just as for Experiment 1, the low average score for Test 1 was again what we had expected, since the 15 words employed in the experiment were all supposed to be least familiar to the subjects.

The pre-test 1 results for Group A and C were exactly the same (1.1), indicating that the starting point was exactly the same for the group that was provided translation only and the one that had both translation and visual data. The mean scores for post-test 1 conducted right after the learning session were also very similar for these two groups. The mean score for Group A in post-test 1 was 24.2, and 24.3 for Group C. The difference observed between the two is only 0.1.

If we shift our attention to the remaining two groups, we find it interesting to see that the mean scores for Group B and D in pre-test 1 were slightly higher than those of either Group A or C, yet, the mean scores for post-test 1 in the former were lower than the latter. This tendency for Group A and C to outscore the remaining two groups persists throughout. In pre-test 2, which was conducted a week later from the learning session, Group C scored the highest (13.8), followed by Group A (13.7), but the difference again was merely 0.1. In the end, all four groups were able to attain fairly good scores, but here again, Group C had the highest mean score followed by Group A, then D, then B, but the difference between Group A and C was slight (0.4). If we exclude the pre-test 1 scores, the result summarized in Table 2 depicts the fact that the learning effect varied greatly as to whether the material used aural data or not. Unlike the result obtained for Experiment 1, however, aural data did not work in favor of enhancing the learning effect. On the contrary, the mean scores for Group B (aur) and Group D (vis + aur), the two groups that employed aural data in the learning material, were constantly lower than the remaining two groups that did not employ aural data.

In addition to the mean scores for each group, we decided to measure the memory retention rates. Table 3 shows the experimental results of the memory retention rates of the four types of learning materials. The memory retention rate here refers to the difference observed between the results for post-test 1 and pre-test 2, conducted one week apart.

We conducted ANOVA based on the result of the memory retention rates, and found the F number for Factor 2 (aural) to be 7.81, as shown in Table 4

Factor 1 (visual)	Without		With	
Factor 2 (aural)	Without (= Group A)	With (= Group B)	Without (= Group C)	With (= Group D)
Data 1	0.333	0.267	0.567	0.133
Data 2	0.233	0.308	0.833	0.357
Data 3	0.233	0.333	0.267	0.607
Data 4	0.545	0.267	0.429	0.517
Data 5	0.367	0.067	0.367	0.375
Data 6	0.444	0.071	0.267	0.300
Data 7	0.367	0.133	0.130	0.300
Data 8	0.533	0.133	0.321	0.214
Data 9	0.733	0.364	0.233	0.000
Data 10	0.600	0.462	0.633	0.100
Number of Data	10	10	10	10
Average	0.439	0.240	0.405	0.290
SD	0.155	0.127	0.203	0.177

Table 3: Learning material and memory retention rates

Factors	Square sums	DOF	Mean squares	F numbers
Factor 1	6.17E-04	1	6.17E-04	1.97E-02
Factor 2	2.45E-01	1	2.45E-01	7.81E+00
Interaction	1.78E-02	1	1.78E-02	5.67E-01
Residual	1.13E+00	36	3.13E-02	
Total	1.39E+00	39		

Table 4: Result of ANOVA

If we put forward the null hypothesis that "there is no difference between the memory retention rates of learning material with sound and without," the result obtained makes it

possible to refute this with a significance level of 0.01. That is, from the average memory retention rates, we can conclude that the learning material without sound is superior to the one with sound. On the other hand, no significant difference could be observed for Factor 1 (visual). Furthermore, there was no significant difference in the interaction of these two factors either.

Since Experiment 2 was a follow-up to Experiment 1, both experiments were conducted under similar conditions. Experiment 2 parallels Experiment 1 in that visual data did not have a particularly positive effect on the participants' learning. The findings obtained from both experiments counter the general claim in the literature that visual data enhance vocabulary learning. However, as we have already seen above, the results of the two experiments came out as completely different in terms of the role that aural data play in vocabulary attainment. In Experiment 1, the material employing aural data only brought about the best result. The result for Experiment 2, however, showed that the two groups using aural data (whether with or without visual data) did not do as well as the groups that did not use aural data. The statistics based on the memory retention rates clearly indicated that aural data actually had a bad effect on vocabulary learning. Furthermore, the group that employed the translation only material actually did quite as well as the group that employed visual data in addition to the translation.

Visual data may be useful, but there are limitations to the type of words that can be expressed visually. Our results indicate that vocabulary attainment may be achieved without relying too much on visual data. If so, then this may shed new light on ways of designing learning material for online use. Instead of trying to force oneself to link text data with visual data or aural data for all lexical items, incorporating flexibility into the vocabulary learning system may be the key for providing a better learning tool. It is generally the case that most of the words that we need to learn in a foreign language are abstract and difficult to express as visual data. If the text only data (translation) can do as well a job as a text (translation) + visual data, then we may not have to worry about finding suitable visual data for these types of abstract words. We, however, do not intend to claim that visual data should be totally denied in vocabulary learning.

It is true that the effectiveness of visual data was not proven in terms of the scores obtained in the experiments. However, the results of the questionnaire that we had conducted on participants of Experiment 2 (right after post-test 2) indicate that there were many who felt that using visual data actually facilitated the learning task. Some participants in Group C (vis) and D (vis + aur) mentioned that the learning task was, in fact, actually fun and entertaining. Another point worth mentioning is that when we were observing the participants during the test session, we found that many participants of Group A (tr) were trying to answer the questions in alphabetical order, that is, they started answering from *crumble* (the first alphabetical word on the list) and ended with *wither* (the last alphabetical word). Since we had presented the vocabulary list in alphabetical order in the pre-tests and the learning process, these participants may merely have been memorizing the words by rote. Since the lexical items on the post-tests were all randomly ordered, they may have been rearranging the words in the order that were presented in the learning process. This tendency to prefer alphabetical order was not observed for the participants in Group C. In other words, almost all of the participants in this group were writing down the answers to the test questions regardless of the order. They may have been able to answer in any order because the way they memorized the words was not by rote (as was the case for the participants in Group A).

If we take this point into consideration, although Group A and C both brought about similar results throughout the two sets of pre-test and post-test that we had conducted, the quality of their understanding of each lexical item may be different. If participants in Group C could answer the test questions regardless of word order, this may mean that they may have achieved a higher level of understanding compared to those in Group A. If we had conducted the post-tests on a longer time span, then we may have been able to measure the difference of memory retention rate between the two groups more clearly. Furthermore, by increasing the number of lexical items, or testing the meaning retention from a different perspective other than translation (e.g. reading comprehension test, definition test, etc.), this difference may further easily be teased out. This is left for future research.

Before ending, we need to consider some of the possible reasons as to why aural data had such a bad effect on the learning effects in Experiment 2. One possible factor may have been the environment under which the experiment was conducted. Although we made best effort to conduct the experiments in quiet a place as possible, there were several occasions during the learning/testing sessions where we could not prepare as suitable an environment as Experiment 1. Both experiments were conducted at different universities, and this may have affected the results. It may be necessary to prepare better controlled environments so that the participants will not be influenced by the surrounding sounds such as passing cars and so on.

6 Concluding remarks

In this paper, we presented the results obtained from the two experiments that tested the learning effects of different types of vocabulary material. By employing the vocabulary creating system that we had developed for mobile devices, we created the following 4 different types of material: 1. translation only, 2. translation + aural data, 3. translation + visual data, 4. translation + aural + visual data. In Experiment 1, the 59 subjects were instructed to memorize the lexical items that were stored in 4 different types just mentioned above, and vocabulary tests were administered in order to measure the memory retention rate. Contrary to claims found in the literature, visual data did not provide much aid in facilitating vocabulary attainment, since the mean test scores were similar between translation only and translation + visual material. A similar tendency could be read off from the results from Experiment 2. This finding has some important implications for language teaching, especially for lexical items that are difficult to relate to visual images, such as abstract words. It is interesting to note that in the second experiment, the learners who employed the translation only material did as well as those who used both translation and visual data. Furthermore, the translation only group outscored the group that employed translation + aural + visual data. Based on this fact, we conclude that it is most likely that vocabulary attainment may be achieved without relying too much on visual data.

Acknowledgments

The authors are grateful to Takeshi Goto and Marie Matsumoto for their contribution to the experiments presented. This study is partly supported by Ministry of Education, Culture, Sports, Science and Technology (MEXT) Fund for Promoting Research on Symbiotic In-

formation Technology. It is also partly supported by the MEXT Special Coordination Funds for Promoting Science and Technology.

Bibliography

ALLEN, V. F. (1983). *Techniques in teaching vocabulary.* Oxford and New York, Oxford University Press.

AL-SEYGHAYAR, K. (2001). "The effect of multimedia annotation modes in L2 acquisition: a comparative." *Language learning and technology, 5*, 202-232.

AMEMIYA, S., HASEGAWA, K., KANEKO, K., MIYAKODA, H, AND TSUKAHARA, W. (2007). "Long-term memory of foreign-word learning by short movies for iPods." *Proc. of the 7th IEEE International conference on advanced learning technologies*, 561-563.

HASEGAWA, K., AMEMIYA, S., ISHIKAWA, M., KANEKO, K., MIYAKODA, H, AND TSUKAHARA, W. (2007). "Promoting autonomous learning: a multilinguistic word learning system based on iPod." *Proc. of the 2007 International conference on ESL/ EFL*, 70-83.

SCHMITT, N., AND MCCARTHY, M. (2005) *Vocabulary: description, acquisition, and pedagogy*. Cambridge, Cambridge University Press.

THORNBURY, S. (2002). *How to teach vocabulary*. London, Longman.

YEH, Y., AND WANG, C. (2003). "Effects of multimedia vocabulary annotations and learning styles on vocabulary learning." *CALICO Journal*, 21 (1), 131-144.

Frank Binder, Harald Lüngen, Henning Lobin

Sprachressourcen in der Lehre – Erfahrungen, Einsatzszenarien, Nutzerwünsche

1 Einleitung

Im zweiten Teil dieses Themenheftes steht eine andere Konstellation im Mittelpunkt. Hier geht es vor allem um die Frage, wie die forschungsbezogene Nutzung von Sprachressourcen (Daten, Werkzeuge, Services) in der akademischen Lehre vermittelt werden kann und welche Rahmenbedingungen den Einsatz von Sprachressourcen in der Lehre erleichtern können.

Vor dem Hintergrund vielfältiger aktueller Bestrebungen, die Möglichkeiten der digitalen Geisteswissenschaften durch die Schaffung von Forschungsinfrastrukturen zu erweitern, stellt sich neben den anvisierten forschungsbezogenen Nutzungsszenarien digitaler Ressourcen auch die Frage, wie entsprechende Inhalte und Arbeitsweisen in die akademische Lehre integriert werden können. Vielerorts werden Sprachressourcen in Forschungsprojekten erarbeitet und genutzt. Neben der langfristigen Sicherung dieser Ressourcen in technischen Infrastrukturen ist deren nachhaltige Nutzung und Pflege im Interesse sowohl der Nutzer als auch der Anbieter. Im Sinne des Humboldt'schen Ideals der Einheit von Forschung und Lehre ist dabei insbesondere der Transfer von der Forschung in die Lehre und zurück von Interesse.

Den Lernenden bzw. den Studierenden erschließen sich die Möglichkeiten der Nutzung von Sprachressourcen und Werkzeugen, und sie werden auf lange Sicht zu deren Nutzern, die wiederum den Fortbestand der Ressourcen sichern. Eine Schlüsselrolle zwischen den Anbietern und den zukünftigen Nutzerinnen und Nutzern kommt dabei den Dozentinnen und Dozenten zu, die die Technik und den forschungsmethodischen Umgang mit den Ressourcen vermitteln. Neben der Arbeit mit akademischen Lehrwerken spielt dabei auch die unmittelbare Nutzung von Sprachressourcen eine zunehmend stärkere Rolle, sei es direkt in den Lehrveranstaltungen oder im Rahmen von Hausarbeiten oder studentischen Projekten. Der Einsatz von Sprachressourcen stellt dabei in mehrfacher Hinsicht eine Herausforderung dar, weist jedoch auch Parallelen zu anderen Bereichen auf, in denen computerbasierte Verfahren in Forschung und Lehre genutzt werden.

Vor diesem Hintergrund wurde im Rahmen des vom BMBF geförderten Projekts D-SPIN[1] (Deutsche Sprachressourcen-Infrastruktur) ein Workshop zu „Sprachressourcen in der Lehre" veranstaltet. Aus einigen Vorträgen des Workshops sind nun Beiträge für dieses Themenheft hervorgegangen. Wir bedanken uns ausdrücklich bei den Autorinnen und Autoren[2], welche sich bereit erklärt haben, die jeweiligen Workshopbeiträge zu Artikeln für dieses Themenheft aufzubereiten.

[1] D-SPIN: Deutsche Sprachressourcen-Infrastruktur: http://d-spin.org

[2] Auf die doppelte Nennung von Personenbezeichnung in maskuliner und femininer Form wird im weiteren Verlauf des Artikels zugunsten der Lesbarkeit stellenweise verzichtet. Männliche Personen-, Berufs- und Rollenbezeichnungen schließen dabei entsprechende weibliche Bezeichnung mit ein.

Die entstandenen Beiträge sind in diesem Teil des Bandes so angeordnet, dass zwei überblicksartige Artikel als Rahmen für drei Beiträge zu spezielleren Themen dienen (Sprachressourcen in Szenarien der theoretischen und der Computerlinguistik, für die historische Semantik und in der Lehrerausbildung).

In diesem dem zweiten Heftteil vorangestellten Artikel stellen wir zum einen die weiteren Beiträge zu diesem Themenschwerpunkt vor und präsentieren zum anderen noch einmal die Gesamtperspektive auf den zugrundeliegenden Workshop.

2 Vorstellung der Beiträge in diesem Teil des Themenheftes

Im ersten Beitrag, *Chancen und Probleme der Nutzung von Korpora, Taggern und anderen Sprachressourcen in Seminaren*, geht Heike Zinsmeister von didaktischen Überlegungen aus und diskutiert zunächst die etablierte Bühnen-Metaphorik zum Einsatz von Sprachressourcen in der Lehre, welche – so stellt die Autorin fest – in der Praxis oftmals um die didaktisch problematische Kategorie des „Schattenspiels" erweitert werden muss. Indem sie anschließend existierende Evaluationskriterien für sprachverarbeitende Software an Korpusressourcen und den Lehrkontext adaptiert, entwickelt die Autorin eine umfangreiche Liste von Entscheidungskriterien für den Einsatz von Sprachressourcen in der Lehre. Kriterien wie z.B. „Sozialform", „Technisches Knowhow", „Lizensierung" und „Operabilität" sind zu Dimensionen wie „Nutzer" und „Verfügbarkeit" gruppiert und in Form von Fragen, die ihre möglichen Ausprägungen spezifizieren, weiter ausgeführt. Im anschließenden Diskussionsteil werden diese möglichen Ausprägungen anhand von Beispielen des Einsatzes von konkreten Ressourcen in Lehreinheiten von computerlinguistischen Seminaren der Autorin ausführlich illustriert, wobei auch unterschiedliche Ausprägungen für Studenten eines computerlinguistischen Studiengangs vs. eines sprachwissenschaftlichen Studiengangs eine Rolle spielen.

In dem Beitrag *Digitale Korpora in der Lehre – Anwendungsbeispiel aus der Theoretischen Linguistik und der Computerlinguistik* leitet Stefanie Dipper Anforderungen an Sprachressourcen aus vier Fallbeispielen, d.h. Lehrsituationen in den Fächern Theoretische Linguistik und Computerlingistik her. Dabei arbeitet sie die unterschiedlichen, teilweise konträren Anforderungen dieser beiden Fächer heraus. So plädiert sie für die Bereitstellung von webbasierten Korpus-Abfragetools und Annotationswerkzeugen für die Lehre in der theoretischen Linguistik, in der die Bildung und Überprüfung von linguistischen Theorien (z.B. Regeln der Vorfeldbesetzung im Deutschen) anhand von Korpusbelegen im Vordergrund steht. Hingegen ist die Lehre in der Computerlinguistik, in der es häufig um die Vermittlung von Methoden der Anwendungs- oder Ressourcenentwicklung geht, auf herunterladbare und lokal installierbare Korpora und Werkzeuge angewiesen, um Daten flexibel vor- und weiterverarbeiten und Werkzeuge in eigene spezifische Verarbeitungsketten einbinden zu können (z.B. für eine Anwendung zur *authorship attribution*).

Die beiden Beiträge von Zinsmeister und Dipper verfolgen somit komplementäre Strategien, indem Zinsmeister top-down und deduktiv Kriterien für Sprachressourcen aus einem didaktischen Konzept und den EAGLES-Evaluationskriterien für Software entwickelt und diese dann auf reale Unterrichtssituationen anwendet. Dipper hingegen beginnt bottom-up und induktiv mit der Beschreibung von vier Unterrichtsszenarien, anhand derer sie abstra-

hierend die unterschiedlichen Anforderungen an Sprachressourcen ableitet und schließlich vergleicht.

Im dritten Beitrag mit dem Titel *Der eHumanities Desktop als Werkzeug in der historischen Semantik: Funktionsspektrum und Einsatzszenarien* schlagen Alexander Mehler et al. eine andere Richtung ein. Mit dem „eHumanities Desktop" stellen Sie ein System vor, welches den integrierten, webbasierten Zugriff auf Sprachressourcen und korpuslinguistische Werkzeuge erlaubt, und illustrieren dessen Einsatzmöglichkeiten für das Forschungsgebiet der Historischen Semantik. Nachdem die Arbeit am „eHumanities Desktop" vor dem Hintergrund der *Digital Humanities* verortet wird, liefern die Autoren zunächst eine kompakte Darstellung des Forschungsgebietes der Historischen Semantik, bevor sie das Funktionsspektrum des „eHumanities Desktop" vorstellen und anschließend auf verschiedene Lehrkontexte eingehen, in denen das System bereits genutzt wurde. Die komplett webbasierte Bereitstellung des Systems erleichtert dabei den Einsatz in der fachwissenschaftlichen Forschung und Lehre, bei dem aus Sicht der Anwender ein Forschungsinteresse im Mittelpunkt steht und die technischen Hürden als solche zu minimieren sind. Erste Erfahrungen im Einsatz in der Lehre an der Universität Frankfurt schildert das Autorenteam anhand verschiedener Veranstaltungsformen, hier Seminare und Übungen sowie Studiengruppen und Anwenderworkshops. In der abschließenden Diskussion werden wiederum verschiedene Zielgruppen identifiziert, welche entsprechend unterschiedliche Anforderungen hinsichtlich der Weiterentwicklung des „eHumanities Desktop" erkennen lassen.

Im vierten Beitrag *Digitale Sprachressourcen in Lehramtsstudiengängen: Kompetenzen – Erfahrungen – Desiderate* stellen Michael Beißwenger und Angelika Storrer ein weitreichendes Einsatzgebiet für Sprachressourcen dar. Vor dem Hintergrund gesamtgesellschaftlicher Entwicklungen wie etwa der generellen Zunahme internetbasierter Kommunikation schildern die Autoren, wie im Bereich der Lehrerausbildung sowohl schulische Lehrpläne als auch Curricula für germanistische Lehramtsstudiengänge inzwischen diverse Anknüpfungspunkte enthalten, die die Arbeit mit digitalen Sprachressourcen in beiden Kontexten motivieren, wenn nicht gar erfordern. Die Autoren unterscheiden dabei zwei Typen didaktischer Konzepte zur Einbindung von Sprachressourcen in die Lehrerausbildung und stellen die von ihnen entsprechend konzipierten und an der TU Dortmund angebotenen Lehrveranstaltungen vor. Beide Lehrveranstaltungen bzw. Konzepttypen werden detailliert erläutert, durch zugehörige Aufgabenbeispiele illustriert und daraus gewonnene Erfahrungen diskutiert. Die Autoren liefern somit wertvolle Anregungen für die Integration der Arbeit mit Sprachressourcen in sprachbezogene Lehramtsstudiengänge. Der Beitrag wird durch eine Auflistung der genutzten Materialien und eingesetzten Ressourcen abgerundet.

In dem letzten Beitrag des Bandes, *Korpuslinguistik in der linguistischen Lehre: Erfolge und Misserfolge* berichtet Noah Bubenhofer im ersten Teil von seinen Erfahrungen als Dozent an den Universitäten Zürich und Mannheim, an denen er Einführungen in die Korpuslinguistik und sprachwissenschaftliche Seminare durchführte. Zunächst wird das didaktische und inhaltliche Konzept der Einführungsveranstaltungen erläutert. Um das Potenzial von Korpuslinguistik in sprachwissenschaftlichen Seminaren zu illustrieren, werden verschiedene Seminar- und Abschlussarbeiten vorgestellt, in denen Forschungsfragen aus der Semantik, der Textlinguistik und der Diskurs- und Kulturanalyse mit korpuslinguistischen

Methoden angegangen wurden. Im zweiten Teil präsentiert Bubenhofer als Autor und Betreuer einer Online-Einführung in die Korpuslinguistik eine Klassifikation der Bedarfe der Nutzer seines Webangebots anhand der Zugriffsstatistiken. Das Fazit fasst die sich aus den beiden Teilen ergebenden Anforderungen an Werkzeuge und Softwaremodule für die Korpuserstellung und -analyse zusammen, die gewährleistet sein müssen, um Studierende mit philologisch-sprachwissenschaftlichem Hintergrund erfolgreich an das empirische Arbeiten heranzuführen.

Wie bereits erwähnt, haben alle fünf Beiträge in diesem Teil des Themenheftes ihren Ursprung in einem gleichnamigen Workshop, zu welchem wir im Folgenden einen Gesamtüberblick darstellen.

3 Hintergrund: Der D-SPIN Workshop zu „Sprachressourcen in der Lehre"

Im Rahmen des Projekts D-SPIN wurde ein Workshop zu *Sprachressourcen in der Lehre: Erfahrungen, Einsatzszenarien, Nutzerwünsche* von den Mitherausgebern Frank Binder, Henning Lobin und Harald Lüngen organisiert. Dieser fand am 18.1.2011 an der Berlin-Brandenburgischen Akademie der Wissenschaften statt. An dem Workshop nahmen ca. 30 Wissenschaftlerinnen und Wissenschaftler teil, zum einen Unversitätsdozenten, die in ihrer Lehre Sprachressourcen einsetzen und zum anderen Mitarbeiter von Einrichtungen, die Sprachressourcen bereit stellen. In acht Vorträgen (siehe Ablaufschema in Abbildung 1) stellten Lehrende ihre Erfahrungen mit dem Einsatz von Sprachressourcen in der universitären Lehre in Studiengängen wie Computerlinguistik, Allgemeine Sprachwissenschaft, Geschichtswissenschaft und Literaturwissenschaft/Philologien dar und formulierten Anforderungen und Wünsche an die Bereitsteller von Ressourcen für die Zukunft

Im Vorfeld des Workshops waren bereits eine Liste von Leitfragen zum Thema bekannt gemacht worden (siehe Abbildung 2). In mehreren Diskussions-Slots ergab sich somit ein anregender Austausch zwischen Lehrenden und Ressourcen-Providern. Im Folgenden liefern wir zunächst eine Gesamtperspektive auf die im Workshop vertretenen Positionen, bevor wir einige Antwortvorschläge zu den Leitfragen wiedergeben.

3.1 Zusammenfassung der Aussagen des Workshops

Insgesamt wurde deutlich, dass Sprachressourcen (Daten, Werkzeuge und Services) in der akademischen Lehre in sehr verschiedenartigen Szenarien und Kontexten eingesetzt werden, da sie in unterschiedlichen curricularen Zusammenhängen eine Rolle spielen. Gleichzeitig sind einige übergeordnete Tendenzen erkennbar, welche wir hier mit einer kurzen Synopsis darstellen möchten. Im Anschluss an diese Darstellung sind einige Antwortvorschläge zu den Leitfragen des Workshops zusammengetragen, so wie sie im Kontext des Workshops sichtbar wurden. Materialien zu den Vorträgen der im Folgenden namentlich aufgeführten Vortragenden sind auf der Webseite zum Workshop verlinkt.[3]

[3] Workshop „Sprachressourcen in der Lehre":
http://www.uni-giessen.de/~g91254/dspin-workshop-lehre/

Ablauf des Workshops:

- Begrüßung und Eröffnung
1. Heike Zinsmeister (Universität Konstanz): Chancen und Probleme der Nutzung von Korpora, Taggern und anderen Sprachressourcen in sprachwissenschaftlichen Seminaren – ein Erfahrungsbericht
2. Noah Bubenhofer (IDS Mannheim): Korpuslinguistik in der linguistischen Lehre: Erfolge und Misserfolge
3. Werner Wegstein (Julius-Maximilians-Universität Würzburg): Sprachressourcen in der Lehre: Erfahrungen aus dem Fachgebiet 'Deutsche Sprachwissenschaft'
4. Alexander Mehler (Goethe-Universität Frankfurt): eHumanities Desktop und historische Semantik
- Moderierte Diskussion (35 min)
5. Cristina Vertan (Universität Hamburg): Sprachressourcen in der Lehre: Erfahrungen aus der historischen Korpuslinguistik
6. Stefanie Dipper (Ruhr-Universität Bochum): Elektronische Korpora in der Lehre - Anwendungsbeispiele aus der theoretischen und der Computerlinguistik
7. Michael Beißwenger / Angelika Storrer (Technische Universität Dortmund): Digitale Sprachressourcen in den Lehramtsstudiengängen: Kompetenzen - Erfahrungen – Desiderate
8. Sabine Bartsch (Technische Universität Darmstadt): Prototypen und Processing Chains: Werkzeuge und Kompetenzen für die linguistische Sprachverarbeitung
- Moderierte Diskussion (45 min)
- Schlussworte, Verabschiedung

Abbildung 1: Vorträge und Ablauf des Workshops „Sprachressourcen in der Lehre“ am 18.01.2011 in Berlin

Betrachtet man zunächst die zugrunde liegenden Daten, also Text- oder Sprachmaterial, so wird deutlich, dass für viele fachwissenschaftliche Fragestellungen – im Unterschied zu methodischen oder verfahrenswissenschaftlichen Fragestellungen – sehr spezifisches Datenmaterial benötigt wird (vgl. Beitrag von Mehler) oder auf renommierte Editionen ganz bestimmter Texte zurückgegriffen werden muss (Wegstein). Große Mengen allgemeinsprachlicher Daten hingegen, wie sie etwa in der theoretischen Linguistik (Dipper), Computerlinguistik (Dipper, Zinsmeister) und Korpuslinguistik (Bubenhofer, Zinsmeister) eine Rolle spielen, dienen einerseits der Entwicklung sprachtechnologischer oder computerlinguistischer Werkzeuge, andererseits vor allem als Vergleichsgrundlage in differenzanalytischen Verfahren. Es besteht also generell Bedarf an standardkonformen und gleichzeitig möglichst flexibel einsetzbaren Referenzressourcen (z.B. Korpora), insbesondere auch für dezidierte Sprachen oder Sprachvarianten. Generell scheint gesprochene Sprache im Rahmen dieses Workshops noch unterrepräsentiert zu sein, ebenso wie multimodale Kombinationen aus Text, Ton, Bild und Video. Es bleibt daher noch zu klären, inwieweit sich die gewonnenen Erkenntnisse aus den Bereichen der primär text-bezogenen Sprachressourcen auch auf diese Bereiche übertragen lassen.

Hinsichtlich der eingesetzten Werkzeuge und Services weist der Workshop in zwei zunächst konträre, aber gleichermaßen interessante Richtungen. Beiden Richtungen ist gemein, dass jeweils Portfolios von spezialisierten Werkzeugen eingesetzt und diese zur Anreiche-

rung von Textmaterial meist zu linguistischen Verarbeitungsketten kombiniert werden (Bartsch, Bubenhofer, Dipper, Mehler, Zinsmeister). Einerseits werden Werkzeuge eingesetzt, die man zwar lokal auf Arbeitsrechnern installieren und administrieren muss (Bartsch, Dipper, Zinsmeister), dafür aber im idealen Falle der Verfügbarkeit als *Open Source*-Software auch selbst modifizieren und neu kombinieren kann (Bartsch). Wenn man im Rahmen von technischen Studiengängen oder Lehrveranstaltungen auch die Entwicklung derartiger Werkzeuge beherrschen will, ist diese Art der Verfügbarkeit bisher die optimale. Auch bietet sich ein solches Ökosystem aus quelloffenen Werkzeugen ideal für interdisziplinäre Team-Arbeit unter Studierenden an (Bartsch, Bubenhofer). Ein anderer Ansatz scheint für technikfernere Fachwissenschaften (wie z.B. Vertan, Wegstein, Dipper) vielversprechender: Webbasierte Anwendungsumgebungen (wie z.B. Mehler), in denen die integrierten Werkzeuge auch ohne technisches Know-How genutzt und darüber hinaus perspektivisch auch Dienste zum kollaborativen Arbeiten sowie zur Sicherung und Versionierung von Daten und Werkzeugen angeboten werden können. Beide Ansätze erachten wir als wertvoll, da einerseits sowohl künftige Werkzeugentwickler gewonnen und trainiert werden müssen, andererseits die Früchte ihrer Arbeit so vielen Anwendern wie möglich zur Verfügung stehen sollen. Inwieweit sich beide Ansätze kombinieren lassen, d.h. inwieweit die webbasierte Arbeitsumgebung gleichermaßen für Nutzer und Entwickler der darin bereitgestellten Werkzeuge interessant ist – etwa indem sie neben intuitiver Bedienung und Kombinierbarkeit auch die Modifikation von Werkzeugen erlaubt und die dafür nötige Verwaltung von referenzierbaren Versionen und Varianten integriert –, bleibt zunächst eine offene Frage.

Neben den bereits angesprochenen Nutzergruppen – textbezogen arbeitende Fachwissenschaftler auf der einen Seite und (zukünftige) Entwickler von computerlinguistischen und sprachtechnologischen Werkzeugen auf der anderen Seite – muss eine dritte Nutzergruppe ebenfalls bedacht werden: Texte und Sprache sind – gemäß ihrer grundlegenden Bedeutung in vielen Berufsfeldern – zentraler Lehrgegenstand in einer Reihe von Schulfächern. Dementsprechend sollten Lehrerinnen und Lehrer im Umgang mit Sprachressourcen geschult, ihre Nutzung in Lehrplänen verankert und die Vermittlung der Nutzungsmöglichkeiten in Curricula für Lehramtsstudiengänge integriert werden (Storrer & Beißwenger). Weiterhin ist anzumerken, dass die Integration der Arbeit mit Sprachressourcen in Curricula auch allgemeine berufsrelevante Fähigkeiten schulen kann, etwa in den Bereichen der Erhebung-, Verwaltung und Auswertung verschiedener Formen von Daten. Dies ist insbesondere der Fall, wenn für diese Zwecke auch die Nutzung weit verbreiteter Standardsoftware (z.B. Techniken der Tabellenkalkulation oder Datenbank-Nutzung) oder grundlegende Techniken der Text- und Sprachtechnologie am Praxisbeispiel demonstriert und vermittelt werden.

3.2 Leitfragen und Antworten des Workshops

Der Workshop orientierte sich an einer zuvor den Vortragenden zur Verfügung gestellten Liste von Leitfragen (s. Abb. 2), die im folgenden in kondensierter, nicht abschließender Form auf der Grundlage der Workshop-Beiträge beantwortet werden sollen:

Leitfragen des Workshops:

1. Welche Themen und Sprachressourcen sind für welche Zielgruppen in der Lehre relevant?
2. Wie ist die Nutzung von Sprachressourcen heute in Curricula integriert?
3. Welche Ressourcen werden bevorzugt und gern praktisch eingesetzt? Aus welchen Gründen? In welchen Szenarien? Lassen sich daraus "best practices" hinsichtlich der Bereitstellung von Ressourcen ableiten?
4. Welche Wünsche und Vorschläge zur Verbesserung und Erleichterung des didaktischen Einsatzes von Sprachressourcen gibt es seitens der Lehrenden?
5. Welche möglichen Hürden und Herausforderungen aus Sicht der Anbieter treten dabei auf?
6. Welche Rahmenbedingungen sind nötig, um den Transfer zwischen Forschung und Lehre im Bereich der Sprachressourcen zu fördern?

Abbildung 2: Leitfragen zum Workshop „Sprachressourcen in der Lehre“

1. *Welche Themen und Sprachressourcen sind für welche Zielgruppen in der Lehre relevant?*
   Für einen Einblick in die Vielfalt der Themen und Zielgruppen verweisen wir auf die verlinkten Präsentationen auf der Webseite des Workshops. Die im Workshop repräsentierten Zielgruppen erstrecken sich von Lehramtsstudierenden für sprachliche Fächer über Studierende der Philologien, Geschichtswissenschaften und Sprachwissenschaften (inkl. Computerlinguistik, theoretische Linguistik, praktische Linguistik, d.h. z.B. Feldforschung und Korpuslinguistik) bis in Bereiche der Informatik.
2. *Wie ist die Nutzung von Sprachressourcen heute in Curricula integriert?*
   Sie ist in den verschiedenen Disziplinen und Wissenschaftsbereichen auf unterschiedliche Weisen integriert. In einigen Bereichen sind „Endnutzer“ an einsatzbereiten Sprachressourcen interessiert, in anderen Bereichen werden der Entwicklungsvorgang und dessen handwerkliche Seite oder dessen theoretische Grundlagen in den Vordergrund gestellt. In wieder anderen Bereichen verschränken sich diese Perspektiven – zum Beispiel in interdisziplinären bzw. fächerübergreifenden studentischen Projekten.
3. *Welche Ressourcen werden bevorzugt und gern praktisch eingesetzt? Aus welchen Gründen? In welchen Szenarien? Lassen sich daraus "best practices" hinsichtlich der Bereitstellung von Ressourcen ableiten?* Antworten bezogen auf die beiden grundlegenden Arten von Ressourcen:
   a. Daten
      i. Teilweise ist sehr spezifisches Datenmaterial wie z.B. ganz bestimmte Texte oder auch renommierte Editionen von Texten von Interesse.
      ii. Referenzressourcen werden sowohl für die „Allgemeinsprache“ als auch für dezidierte Sprachen oder Sprachvarianten als Vergleichsgrundlage für differenzanalytische Verfahren benötigt.

b. Werkzeuge
    i. Portfolios spezialisierter Werkzeuge, verknüpft zu linguistischen Verarbeitungsketten, werden disziplinübergreifend genutzt.
    ii. Konkret wurde der Wunsch nach einer expliziten Verknüpfung von Annotationswerkzeugen mit Recherche-Tools für die Abfrage (Querying) von annotierten Daten deutlich.
    iii. Modifizierbarkeit, Anpassbarkeit, Transparenz und Flexibilität sollten gegeben sein, z.B. in einem Ökosystem quelloffener Werkzeuge.
    iv. Weiterhin sollten diese Werkzeuge als integrierte Komponenten einer nutzerfreundlichen webbasierten Arbeitsumgebung bereitgestellt werden.
    v. Die Anknüpfung der Werkzeuge an Standardsoftware zur Erhebung, Verwaltung und Auswertung von Daten (z.B. Tabellenkalkulation, Statistik-Software, Datenbanken) bietet neben der Vermittlung text- und sprachtechnologischer Verfahren und Techniken eine breite Anknüpfungsmöglichkeit in Richtung allgemeiner berufsrelevanter Kompetenzbildung.

4. *Welche Wünsche und Vorschläge zur Verbesserung und Erleichterung des didaktischen Einsatzes von Sprachressourcen gibt es seitens der Lehrenden?*
    a. Daten sollten standardkonform in Formaten vorliegen, die von den diversen Werkzeugen ohne eigene Vorarbeit verarbeitet werden können.
    b. Werkzeuge: An dieser Stelle verweisen wir auf die Vorschläge aus 3.b.ii bis 3.b.iv.
    c. Mehrmals wurde der Wunsch nach einer besseren Dokumentation von Ressourcen benannt, idealerweise auch im Hinblick auf den Einsatz in Lehrsituationen, z.B. durch Tutorials oder E-Learning-Einheiten oder auch durch eine Sammlung von Referenzen auf Studien, in denen die Ressource schon verwendet wurde.
    d. Es gibt einen Bedarf an institutionalisierten Dienstleistungen für Ressourcennutzer wie z.B. Help-Desk-Services und die Bildung von Expertennetzwerken.
5. *Welche möglichen Hürden und Herausforderungen aus Sicht der Anbieter treten dabei auf?*

    Im Kontext von D-SPIN sind diese folgenden Herausforderungen sichtbar:
    a. Es braucht Zeit, Geld, Geduld und Engagement sowie ein gutes Erwartungs-Management.
    b. Die teils aufwändige Klärung rechtlicher Fragestellungen, gegebenenfalls verbunden mit nötigen Maßnahmen des Interessensausgleichs, sowie bestehende Vereinbarungen mit den „Spendern“ von linguistischem Datenmaterial stehen ambitionierten Wünschen teils schlicht entgegen. Die Änderung von Rahmenbedingungen sind nur in engen Grenzen möglich, und

die Achtung bestehender vertraglicher Vereinbarungen ist für die Vertrauensbildung und Reputation essenziell.

c. Die technische Unterstützung rechtlich sicherer und ethisch unbedenklicher Bereitstellungsmöglichkeiten von Sprachressourcen ist ein Feld aktueller und künftiger Forschung und Entwicklung – daher verweisen wir hier noch einmal auf Punkt 5.a

6. *Welche Rahmenbedingungen sind nötig, um den Transfer zwischen Forschung und Lehre im Bereich der Sprachressourcen zu fördern?*

Zwei Anregungen traten zutage:

a. Die Integration der Nutzung von Sprachressourcen in Curricula und Lehrpläne sollte vorangetrieben werden.

b. Implizit: Technisch-administrative und Community-bildende Arbeit im Bereich der Aufbereitung und Verfügbarmachung von Sprachressourcen sollten weiterhin anerkannt und gefördert werden.

3.3 Training und Ausbildung im Kontext von CLARIN/D-SPIN

Der in diesem Beitrag dargestellte Workshop zu „Sprachressourcen in der Lehre“ ist Teil einer Reihe von Aktivitäten im Bereich Training und Ausbildung im Kontext der Forschungsinfrastrukturprojekte CLARIN[4]/D-SPIN während ihrer „Preparatory Phase“ 2008-2011 (s. ESFRI 2011). Diese reichen von initialen Orientierungaktivitäten über verschiedene durchgeführte Lehrveranstaltungen bis hin zur Bereitstellung von Tutorials.

In einer ersten Phase wurden innerhalb der Projektkonsortien auf nationaler und europäischer Ebene Ideen und Strategien für die Arbeit in den einzelnen Bereichen zusammengetragen und diskutiert, so auch für den Bereich Training und Ausbildung (siehe etwa BINDER & CRISTEA 2009). Strategische Leitfragen zu den Aktivitäten in diesem Bereich wurden in dieser Zeit gesammelt und im Projektverlauf dokumentiert (AHLBORN & BINDER 2010). So wurde deutlich, dass man vielfältige Angebote benötigt und zwischen regionalen, nationalen und internationalen Angeboten sowie zwischen universitären und außercurricularen Angeboten für verschiedene Qualifikationsstufen unterscheiden muss, zusätzlich zur jeweiligen thematischen Schwerpunktsetzung. Weiterhin besteht Bedarf an Training und Ausbildung sowohl auf Seiten der teils stark spezialisierten Nutzer als auch auf Seiten der Entwicklerteams und Anbieter von Sprachressourcen. Ebenfalls in dieser Phase wurde durch den „Sprachressourcen-Gipfel“ in Mannheim 2009 frühzeitig Kontakt zu verschiedenen Fachcommunities aufgebaut und die Vielfalt der vorhandenen Ressourcen und Interessensgebiete verdeutlicht.

In der folgenden Zeit verfolgte das D-SPIN Arbeitspaket 6 „Training und Ausbildung“ vor allem eine nutzerorientierte Strategie. Neben der Integration des Themengebietes in lokale universitäre Curricula geisteswissenschaftlicher Studiengänge war vor allem die Ausrichtung der D-SPIN Sommerschule 2010 ein Höhepunkt in dieser Phase des Projekts (BINDER ET AL. 2010).

[4] CLARIN: Common Language Resources and Technology Infrastructure: http://www.clarin.eu

Für die Zielgruppe der Entwickler und Anbieter von Sprachressourcen wurden durch die Projektpartner parallel weitere Angebote geschaffen. So werden derzeit und auch in Zukunft wiederholt Workshops und Trainingskurse zu technischen Themen angeboten – einmal im Zuge des Überganges zur neuen Projektphase in CLARIN-D, aber auch im Rahmen von CLARA[5] – einem europäischen „Initial Training Network", welches thematisch und institutionell eng mit CLARIN verknüpft ist. Für WebLicht, der in D-SPIN gemeinsam von Teams aus Tübingen, Leipzig, Berlin, Stuttgart und Mannheim entwickelten Web-Anwendung für linguistische Verarbeitungsketten (HINRICHS ET AL. 2009, HINRICHS ET AL. 2010), stehen auf der Webseite verschiedene Tutorials unter anderem zur Integration eigener linguistischer WebServices bereit.[6]

Spätestens auf der gemeinsam von CLARIN und DARIAH[7] ausgerichteten Konferenz „Supporting the Digital Humanities" (SDH) im Oktober 2010 in Wien wurde deutlich, dass nutzerorientierte Ausbildung und Lehre durch die beteiligten Partner der entstehenden Forschungsinfrastrukturen ein bedeutender Erfolgsfaktor sind und auch zukünftig gemeinsame Anstrengungen erfordern.

4 Diskussion und Ausblick

Mit dem Übergang von der „Preparatory Phase" zur „Implementation Phase"[8] des europäischen Projekts CLARIN endet auf nationaler Ebene das Projekt D-SPIN, vor dessen Hintergrund der Workshop zu „Sprachressourcen in der Lehre" konzipiert wurde, und das unmittelbar anschließende Verbundprojekt CLARIN-D nimmt seine Arbeit auf. Die Fragen nach Möglichkeiten einer besseren Integration der Nutzung von elektronischen Sprachressourcen in die akademische Lehre bleiben damit auf der Agenda.

Mit der Konzeption des Workshops zu „Sprachressourcen in der Lehre" war es beabsichtigt, den Status Quo zur Nutzung von Sprachressourcen in der Lehre zu erkunden, von dem aus es nun gilt, die nächsten Schritte zu identifizieren und anzugehen. Im Idealfall werden so Lücken im Vorhandenen identifiziert, deren Schließung eine Erleichterung für die Nutzung von Sprachressourcen darstellt. Wenn also etwa etablierte Werkzeuge eine an aktuelle Betriebssysteme angepasste Benutzungsschnittstelle – sei diese nun webbasiert oder nicht – erhalten können, andere Ressourcen unter neuen, weniger restriktiven Lizenzen bereitgestellt werden können oder an verschiedenen Stellen die Zugänglichkeit zu und Interoperabilität zwischen Werkzeugen und Daten verbessert werden kann, dann sind dies zweifelsohne hilfreiche Entwicklungen. Dass derartige Anforderungen insbesondere auch aus den Bereichen der akademischen Lehre kommen können, zeigt deren Relevanz als Einsatzfeld von Sprachressourcen. Gleichfalls bietet das Konzept eines solchen Workshops stets Raum zur Weiterentwicklung, den wir im Folgenden ein wenig skizzieren wollen.

Der Fokus des Workshops und seiner hier veröffentlichten Beiträge liegt stark auf textuellen Ressourcen. Wenngleich die sich dahinter verbergende Vielfalt nicht unterschätzt

[5] CLARA: Common Language Resources and their Applications – a Marie Curie ITN. https://clara.uib.no/
[6] WebLicht Tutorials: http://weblicht.sfs.uni-tuebingen.de/englisch/tutorials/html/index.html
[7] DARIAH: Digital Research Infrastructure for the Arts and Humanities: http://dariah.eu/
[8] Die Bezeichnungen der einzelnen Konstruktionsphasen europäischer Forschungs-infrastrukturen wurden wie in ESFRI (2011) verwendet.

werden sollte, scheint es doch geboten, diesen Fokus auszuweiten oder in ähnlichen Formaten multimodale Ressourcen in den Vordergrund zu stellen. Dabei wären sowohl Kombinationen aus Texten und Bildern zu betrachten als auch der Umgang mit Sammlungen von Audio- und Video-Daten, welche zum Zwecke wissenschaftlicher Analysen üblicherweise textuell transkribiert und anschließend mit weiteren Annotationen versehen werden. Auch für derartige Ressourcen bieten die aktuellen Infrastrukturentwicklungen neue Perspektiven, die es in die akademische (Methoden-)Lehre zu integrieren gilt.

Ebenfalls deutlich wird eine sprachwissenschaftliche Gewichtung der Beiträge im Workshop und in diesem Themenheft, während CLARIN/D-SPIN einen größeren Kreis potentieller Nutzer erreichen will. Über die Zusammenarbeit mit diversen Fachcommunities soll diese Fokussierung im Verlauf der derzeit beginnenden Konstruktionsphase von CLARIN-D ausgeweitet werden. Parallel dazu ist es ebenso wichtig, neben der Ausrichtung auf fachwissenschaftliche Nutzer, auch für Entwickler und Anbieter von Sprachressourcen entsprechende Themen und Lehrveranstaltungen in technisch orientierte Studiengänge zu integrieren bzw. vorhandene Angebote zu pflegen.

Die Konzeption des Workshops ist weiterhin gezeichnet durch seine Infrastrukturperspektive, aus der die Lehre als ein Einsatzgebiet von Sprachressourcen erscheint. In dieser Konstellation denkt man leicht eher in Begriffen wie „Einsatzszenarien“, als sich etwa über „Lernziele“ Gedanken zu machen. Dabei ist die Didaktisierung eine Herausforderung an sich. Aus Sicht der Hochschullehrerinnen und Hochschullehrer ist der didaktische Nutzen bzw. Mehrwert, welcher mit dem Einsatz von Sprachressourcen verbunden sein kann, nicht als selbstverständlich anzusehen. Erfahrungen aus anderen Bereichen, etwa der softwaregestützten Statistik in empirischen Geistes- und Sozialwissenschaften oder des computergestützten Fremdsprachenlernens, zeigen, dass man im Spannungsfeld zwischen Computereinsatz und didaktischen Zielen gleichermaßen Erfolge und Kontroversen finden kann. Es sollte sich daher lohnen, die Entwicklungen um den Einsatz computerbasierter Sprachressourcen in fachwissenschaftlichen Lehrkontexten auch unter dem Gesichtspunkt der Didaktisierung weiter zu verfolgen. Diesbezüglich ist auch anzumerken, dass im Rahmen des Workshops nicht explizit nach vorhandenen und eingesetzten Lehrwerken – etwa zur Methodenlehre – gefragt wurde, wenngleich natürlich einzelne Beiträge solche Werke benennen. Hier könnte man ansetzen und noch einmal systematisch das derzeitige Angebot betrachten und gegebenenfalls entsprechenden Bedarf ermitteln.

Der Workshop war bewusst auf die deutschsprachige Fachcommunity ausgerichtet. Man kann ihn dadurch einordnen in umfassendere Bestrebungen mit dem Ziel, Inhalte der *Digital Humanities* in die Curricula deutscher Hochschulen einzubinden und diese Entwicklungen sichtbar zu machen (SAHLE 2010). Natürlich gibt es im vorliegenden Themenfeld auch im internationalen Kontext Aktivitäten, durch die der Austausch von Szenarien, Erfahrungsberichten und Empfehlungen zur akademischen Lehre in den *Digital Humanities* innerhalb der Fachcommunities gefördert werden soll (s. etwa HIRSCH 2010 oder TOMASEK & DAVIS 2011 sowie die darin enthaltenen Verweise). Szenarien, die Lehre und Forschung etwa im Bereich kollaborativer Annotation von Korpusdaten eng verzahnen – zum Beispiel realisiert als „’Classroom’ production method“ (BAMMAN & CRANE 2010) –, werden erst durch geeignete Korpusinfrastrukturen ermöglicht. Hier geht es also nicht nur darum, bisherige Lehr- und

Lernformen effizient zu unterstützen, sondern auch neue Formate in der Lehre zu ermöglichen. Derartige Weiterentwicklungen zählen sicherlich zu den wichtigsten Impulsen, die geisteswissenschaftliche Forschungsinfrastrukturen den eHumanities geben können.

Dabei wird CLARIN nicht all diese Gebiete aus sich selbst heraus bedienen können. Vielmehr sind einzelne Fachcommunities im Rahmen der Kooperation mit CLARIN-D, DARIAH-DE und ähnlichen Initiativen angehalten, kompatible Lösungen unter Nutzung der entstehenden Infrastruktur selbst zu gestalten und zu erschaffen. Dies schließt auch den Transfer in die Lehre ein, sei es durch erhöhte Sichtbarkeit für einschlägige studentische Projekte, die Erstellung von praktischem Übungsmaterial oder Beiträgen zu Methodenlehrwerken oder durch regelmäßige Workshops und Schulungen für Forschende und Lehrende zu den Einsatzmöglichkeiten und –erfahrungen der neuen elektronischen Werkzeuge in speziellen fachwissenschaftlichen Kontexten.

Danksagungen

Konzeption, Organisation, Durchführung und Nachbereitung des Workshops „Sprachressourcen in der Lehre“ am 18.01.2011 an der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) wurden ermöglicht durch die Förderung des Bundesministeriums für Bildung und Forschung (BMBF) im Rahmen des Projekts D-SPIN, ergänzt durch Förderungen des Hessischen Ministerium für Wissenschaft und Kunst (HMWK) im Rahmen des LOEWE-Schwerpunkts „Kulturtechniken und ihre Medialisierung“[9].

Für das sorgfältige Reviewing der Beiträge in diesem Teil des Themenheftes danken wir Volker Boehlke, Gertrud Faaß, Gerhard Heyer, Marc Kupietz, Lothar Lemnitzer und Elke Teich. Weiterhin danken wir Heike Zinsmeister für Kommentare zu einer früheren Version des Abschnitts „Leitfragen und Antworten des Workshops“. Besonderer Dank gilt allen Vortragenden und Teilnehmenden des Workshops für die rege Beteiligung und die interessanten Diskussionen.

Literatur

AHLBORN, SVETLANA & FRANK BINDER. (2010). D-SPIN Report R6.1: Training materials and guidelines for their applications. http://weblicht.sfs.uni-tuebingen.de/Reports/D-SPIN_R6.1.pdf

BAMMAN, DAVID & GREGORY CRANE (2010). Corpus Linguistics, Treebanks and the Reinvention of Philology. In Fähnrich, K.-P. and Franczyk, B (Hrsg.) Proceedings of INFORMATIK 2010: Service Science, September 27 - October 01, 2010, Leipzig, volume 2 of Lecture Notes in Informatics, pages 542-551. GI.

BINDER, FRANK & DAN CRISTEA (2009). Training and dissemination in CLARIN – the potential to bridge gaps. In: CLARIN Newsletter #6 June 2009. http://www.clarin.eu/newsletter

BINDER, FRANK, SVETLANA AHLBORN, JOST GIPPERT AND HENNING LOBIN (2010) Connecting with User Communities: D-SPIN Summer School 2010. Poster at SDH+NEERI 2010. Vienna, Austria. http://www.dspin-sommerschule.de/

[9] LOEWE-Schwerpunkt „Kulturtechniken und ihre Medialisierung“: http://www.kulturtechniken.info/

ESFRI (EUROPEAN STRATEGY FORUM ON RESEARCH INFRASTRUCTURES) (2011). Strategy Report on Research Infrastructures - Roadmap 2010. Luxembourg: Publications Office of the European Union. http://dx.doi.org/10.2777/23127

HINRICHS, ERHARD, MARIE HINRICHS, THOMAS ZASTROW, GERHARD HEYER, VOLKER BOEHLKE, UWE QUASTHOFF, HELMUT SCHMID, ULRICH HEID, FABIENNE FRITZINGER, ALEXANDER SIEBERT AND JÖRG DIDAKOWSKI (2009). WebLicht - Web-based LRT services for German. In: Elke Teich, Andreas Witt and Peter Frankhauser (Eds.) GSCL Workshop: Linguistic Processing Pipelines – Book of Abstracts (pp. 11-14)

HINRICHS, MARIE, THOMAS ZASTROW AND ERHARD HINRICHS (2010). WebLicht: Web-based LRT Services in a Distributed eScience Infrastructure. In: Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner and Daniel Tapias (Eds.) Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)

HIRSCH, BRETT D. & MEAGAN TIMNEY (2010). The Importance of Pedagogy: Towards a Companion to Teaching Digital Humanities. Poster at Digital Humanities 2010. London, UK. http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/html/ab-853.html

SAHLE, PATRICK (2010). Digitale Geisteswissenschaften: Studieren! In: Silke Schomburg, Claus Leggewie, Henning Lobin und Cornelius Puschmann (Hrsg.) Beiträge der Tagung „Digitale Wissenschaft - Stand und Entwicklung digital vernetzter Forschung in Deutschland“. (S. 51-56) http://www.hbz-nrw.de/dokumentencenter/veroeffentlichungen/Tagung_Digitale_Wissenschaft.pdf

TOMASEK, KATHRYN & REBECCA FROST DAVIS (2011). Integrating Digital Humanities Projects into the Undergraduate Curriculum. Pre-Conference Workshop at Digital Humanities 2011. Stanford, CA. https://dh2011.stanford.edu/?page_id=499

Heike Zinsmeister

Chancen und Probleme der Nutzung von Korpora, Taggern und anderen Sprachressourcen in Seminaren

1 Einleitung

Mit Korpora oder lexikalisch-semantischen Ressourcen zu arbeiten und dabei Programme zur Aufbereitung oder Analyse der Daten zu nutzen, gehört zum Alltag vieler Computerlinguisten. Computerlinguistische Studiengänge sollten daher ihren Studierenden nicht nur Wissen über Theorien und Algorithmen vermitteln – und eigene Programmierkenntnisse — sondern sie auch auf den Umgang mit vorhandenen Sprachressourcen vorbereiten. Hierbei sind nicht Ressourcen für das *E-Learning* gemeint, sondern Sprachressourcen, die unabhängig von einer Verwendung in der Lehre entwickelt wurden.

Beispiele für solche Sprachressourcen sind:

- Korpora wie das wortartengetaggte und diachron aufbereitete Kernkorpus des DWDS[1] oder die syntaktisch und koreferenzannotierte Tübinger Baumbank Deutsch / Zeitungssprache (TüBa-D/Z)[2].
- Lexikalisch-semantische Ressourcen wie GermaNet[3] oder FrameNet[4].
- Tagsets und Annotationsrichtlinien wie das Stuttgart-Tübingen-Tagset STTS (Schiller et al., 1999) für die Annotation von Wortarten im Deutschen oder die MATE-Annotationsrichtlinien zur Annotation von Koreferenz (Poesio, 2000).
- Programme für die manuelle Annotation wie EXMARaLDA ('EXtensible MARkup Language for Discourse Annotation', (Schmidt, 2004)) für die Mehrebenenannotation von multimodalen Korpora oder MMAX2 ('Multi-Modal Annotation in XML', (Müller and Strube, 2006)).
- Programme zur automatischen Annotation wie der TreeTagger (Schmid, 1997) für Wortartenannotation und syntaktisches Chunking oder der Stanfordparser (Rafferty and Manning, 2008) für syntaktische Konstituenten- und Dependenzannotation.
- Programme zu Indexierung, Recherche und Visualisierung wie die Corpus Workbench mit dem mächtigen Abfrageprogramm Corpus Query Processor CQP[5],

[1] DWDS-Korpora: `http://www.dwds.de/` [Letzter Aufruf der zitierten URLs: 23.03.2011].
[2] TüBa-D/Z: `http://www.sfs.uni-tuebingen.de/tuebadz.shtml`.
[3] GermaNet: `http://www.sfs.uni-tuebingen.de/GermaNet/`.
[4] FrameNet: `http://framenet.icsi.berkeley.edu/`.
[5] CQP: `http://cwb.sourceforge.net/index.php`.

ANNIS2[6] für die Abfrage und Darstellung von mehrebenenannotierten und multimodalen Daten oder TigerSearch[7] für die Abfrage syntaktisch annotierter Korpora.

In der Praxis sind Lehrende beim aktiven Einsatz von Sprachressourcen in Seminaren mit besonderen Anforderungen konfrontiert, die die Durchführung oftmals behindern oder sogar zum Scheitern verurteilen. Der vorliegende Beitrag stellt eine Art Checkliste für die Vorbereitung des Einsatzes von Ressourcen zusammen und diskutiert Probleme, die in Bezug auf einzelne Kriterien der Checkliste auftreten können. Abschließend werden aus dieser Diskussion Wünsche abgeleitet, wie Sprachressourcen bereitgestellt werden sollten, so dass sie in der Lehre leichter einsetzbar werden.

2 Didaktische Überlegungen

2.1 Zielgruppen

Stimmt es wirklich, dass der Einsatz von Sprachressourcen in der Lehre andere Anforderungen mit sich bringt, als die Verwendung von Sprachressourcen in Forschungsprojekten? In einem Szenario, in dem die Studierenden einen fundierten computationellen Hintergrund haben, ist der Unterschied vielleicht nicht groß. Diese Situation ist aber oft nicht gegeben, selbst in grundständigen Studiengängen der Computerlinguistik. Hinzu kommt, dass Seminare mit computerlinguistischen Inhalten nicht nur in den ausgewiesenen Studiengängen vertreten sind, sondern Bestandteil vieler sprachwissenschaftlicher und texttechnologischer Studiengänge sind. Zum Beispiel bietet der Bachelorstudiengang 'Sprachwissenschaft' an der Universität Konstanz[8] in Modul 4 unter 'Weiterführende Gebiete der Linguistik' neben Spracherwerb, Typologie und anderen Gebieten auch ein Seminar 'Computerlinguistik' an. Das Seminar führt Grundideen der Computerlinguistik ein und skizziert die Funktionsweisen von Applikationen wie Wortarten-Tagging, Maschineller Übersetzung, automatischer Grammatik- und Rechtschreibprüfung oder automatischer Textzusammenfassung. Die Studierenden erhalten dabei einen ersten Einblick in die Möglichkeiten (und Grenzen) der maschinellen Sprachverarbeitung. Interessierte Studierende können Computerlinguistik im Masterprogramm 'Speech and Language Processing' vertiefen, in dem Computerlinguistik einen der beiden möglichen Schwerpunkte bildet. Dort ist auch vorgesehen, dass die Studierende eine Programmiersprache erlernen. Dies gilt nicht für das Bachelorangebot.

2.2 Einsatzszenarien

Es sind drei allgemeine Szenarien für den Einsatz von Sprachressourcen in der Lehre denkbar[9]: 'Im Rampenlicht', 'Hinter den Kulissen' und als 'Schattenspiel'. Im er-

[6] ANNIS2: `http://www.sfb632.uni-potsdam.de/d1/annis/`.
[7] TigerSearch: `http://www.wolfganglezius.de/doku.php?id=cl:tigersearch`.
[8] url: `http://ling.uni-konstanz.de/pages/allgemein/bachelor.html`.
[9] Die hier verwendete Bühnen-Metaphorik geht auf Guy Aston zurück, der für den Einsatz von Korpora in der Fremdsprachendidaktik eine 'on stage'-Nutzung von einer 'behind the scene'-Nutzung unterscheidet (vgl. Aston (2000)).

sten Szenario stehen die Ressourcen dahingehend 'im Rampenlicht', dass sie von den Studierenden selbst aktiv im Seminar oder in Übungen genutzt werden. 'Hinter den Kulissen' werden Ressourcen verwendet, wenn die Lehrenden sie für die Erstellung ihrer Lehrmaterialien verwenden, die Nutzung selbst aber in der Lehre nicht thematisieren. Ein Beispiel wäre, die TüBa-D/Z auszuwerten, um interessante Beispiele für eine Seminarstunde zu indirekten Anaphern ('Bridging') zu finden. Ein anderes Beispiel wäre, den TreeTagger auf unterschiedlichen Trainingsdaten zu trainieren, um realistische Zahlen für eine Übungsaufgabe zur Evaluierung von Wortartentaggern zu erhalten. Das dritte Szenario ähnelt einem 'Schattenspiel'. Die Ressourcen werden im Unterricht zwar thematisiert und, zum Beispiel, auf Folien im Unterrichtsvortrag vorgestellt, aber von den Studierenden selbst nicht aktiv genutzt – auch nicht in seminarbegleitenden Übungen. Die Studierenden erhalten dabei ein rein passives Wissen über die Ressourcen.

Untersuchungen zur Lernpsychologie zeigen, dass aktives Lernen den Lernprozess am Besten unterstützt (vgl. z.B. Winteler (2004), Kapitel 10). Umgangssprachlich ausgedrückt, man lernt etwas am Besten, indem man es *tut*. 'Begreifen' hat sehr viel mit 'Greifen' d.h. mit aktivem Handeln zu tun. Aus didaktischer Sicht ist die aktive Verwendung von Sprachressourcen im 'Rampenlicht' daher einer passiven Vorstellung in 'Schattenspielen' vorzuziehen.

2.3 Lernziele

Der Einsatz von Sprachressourcen in der Lehre kann zweierlei Lernzielen dienen. Zum einen kann das Ziel eine *Methoden- bzw. Ressourcenkompetenz* sein, die die Lernenden in die Lage versetzt, die verwendeten Ressourcen selbstständig für eigene Zwecke einzusetzen. Zum Beispiel, mit welchen Kommandozeilen-Befehlen der TreeTagger auf einem Korpus trainiert werden kann, oder mit welcher Maus- und Tastenkombination in MMAX2 eine Koreferenzrelation zwischen zwei 'Markables' markiert wird bzw. welches linguistische Phänomen in der TüBa-D/Z als 'Bridging' annotiert ist. Zum anderen kann die Ressource nur als *Mittel zum Zweck* eingesetzt werden, wenn das eigentliche Lernziel das Verstehen einer (computer-)linguistischen Fragestellung ist. Zum Beispiel kann der TreeTagger in einem Szenario eingesetzt werden, in dem das eigentliche Lernziel die Evaluierung von statistischen Programmen ist, und die Studierenden Kompetenzen im Umgang mit Akkuratheitsbestimmung, Kreuzvalidierung und Konfusionsmatrizen erwerben sollen. Wenn das eigentliche Lernziel Koreferenz und Anaphorik im Deutschen ist, können Studierende in den anaphorischen Annotationen der TüBa-D/Z interessante Beispiele mit Kontext finden und mit dem Programm MMAX2 das Phänomen mit Hilfe von manueller Annotation kennenlernen (vgl. Zinsmeister (2011)).

In der realen Umsetzung findet man Mischformen der beiden Lernziele. Das scheint auch sinnvoll zu sein, weil die eigene Erfahrung zeigt, dass die Studierenden praktische Hürden besser tolerieren, wenn das Lernziel ausdrücklich (auch) die Ressourcenkompetenz selbst ist. Soll die Ressource nur Mittel zum Zweck sein, wird eher ein reibungsloser Ablauf erwartet und nur wenig Aufwand für das Erlernen der Handhabung akzeptiert.

3 Kriterien für den Einsatz von Sprachressourcen 'im Rampenlicht'

Dieser Abschnitt stellt eine Liste von Kriterien für die Nutzung von Sprachressourcen in der Lehre vor. Die Zusammenstellung ist als eine Art Checkliste aufgebaut. Mögliche Ausprägungen der Kriterien werden in Fragen formuliert. Die Kriterien ergeben sich aus der Art des Lernziels und (impliziten) Modellen der Rahmenbedingungen, des Nutzers und der Ressource selbst. Das Ressourcen-Modell beschreibt externe Merkmale der Ressource sowie deren Eigenschaften in konkreten Verwendungskontexten.

Die Auswahl der Kriterien orientiert sich an Evaluationskriterien für sprachverarbeitende Software (vgl. EAGLES (1996), basierend auf dem allgemeinen Qualitätsmodell für Software ISO 9126)[10] und ist um Kriterien für Korpusressourcen und für die Implementierung im Lehrkontext erweitert. Ein Anspruch auf Vollständigkeit besteht nicht.

3.1 Checkliste

Lernziel

- Ist die Kenntnis der Ressource und ihrer Nutzung das Lernziel oder nur ein Mittel zum Zweck?

Rahmenbedingungen

- Zeit: Wie viel Zeit steht für die Einarbeitung und Nutzung der Ressource sowie die Auswertung der Ergebnisse zur Verfügung?
  Soll die Ressource nur innerhalb einer Seminarsitzung genutzt werden, seminarbegleitend (während mehrerer Sitzungen) oder zusätzlich zur Lehrveranstaltung in Seminar- oder Studienarbeiten?

- Ort: An welchem Ort soll die Ressource genutzt werden?
  Im Universitäts-Pool, auf privaten Rechnern an der Universität oder auf privaten Rechnern zuhause?

- Hardware: Welche Hardware steht zur Verfügung?
  Eine homogene Plattform auf Poolrechnern oder eine Vielzahl von Betriebssystemen auf den Privatrechnern der Studierenden?

- Sozialform: In welcher sozialen Form soll die Ressource genutzt werden?
  In Einzelarbeit oder in Gruppenarbeit?

[10] Vielen Dank an Stefanie Dipper, die mich auf diese Qualitätskriterien hinwies, siehe auch Dipper et al. (2004).

Nutzer

- Motivation: Kennen die Studierenden Forschungsfragen, die einen Einsatz der Ressource motivieren? Sind die Studierenden durch eigene (Forschungs-)Interessen motiviert, die Ressource kennenzulernen? Wissen die Studierenden um die Berufsrelevanz der Ressource?
- Technisches Know-How: Sind die Studierenden mit der zu nutzenden Hardware vertraut? Sind die Studierenden bereits mit der Ressource selbst vertraut? Sind die Studierenden mit ähnliche Ressourcen vertraut? Besitzen die Studierenden Programmierkenntnisse (z.B. um die Ressource zu installieren und zu bedienen)?

Verfügbarkeit

- Bereitstellung: Kann man die Ressource online nutzen? Kann die Ressource heruntergeladen werden? Ist die Prozedur, wie man die Ressource erhalten kann transparent?
- Lizenzierung: Muss die Ressource lizenziert werden? Gibt es Gruppen- oder nur Einzellizenzen?
- Kosten: Ist die Ressource für die Lehre kostenfrei?
- Externer Support: Erhalten die Nutzer der Ressource Unterstützung von den Entwicklern oder anderen Experten?

Funktionalität

- Tauglichkeit /Angemessenheit: Bietet die Ressource für das Lernziel relevante Merkmale oder Inhalte bzw. erzeugt sie relevante Analysen?
- Vertrautheit: Entspricht die Ressource in ihrer Funktionalität bekannten Standards oder Konventionen (z.B., dass Copy&Paste mit bekannten Tastenkombinationen möglich ist oder dass bei der Darstellung von Dependenzannotationen die Pfeilspitze einer Kante auf das jeweilige Regens zeigt).

Verwendbarkeit

- Verständlichkeit: Ist der Aufbau der Ressource intuitiv und leicht verständlich? Ist die Ressource gut dokumentiert? Sind ggfs. Eingabe- und Ausgabeformate bekannt?
- Lernbarkeit: Kann die Verwendung der Ressource leicht erlernt werden? Gibt es eine Lernanleitung?
- Operabilität: Ist die Ressource leicht zu handhaben? Zum Beispiel, gibt es intuitive Hilfefunktionen?
- Attraktivität: Macht es Spaß die Ressource zu nutzen?

Effizienz

- Zeit: Wie lang dauert die Nutzung / Ausführung des Programms (z.B. Dauer der Auswertung einzelner Korpussuchanfragen oder Trainingszeiten für ein statistisches Übersetzungsmodell).
- Ressourcennutzung: Beansprucht die Ressource große Rechnerkapazitäten? Benötigt sie Zugriff auf andere Ressourcen?

Portabilität

- Adaptierbarkeit: Kann die Ressource auf verschiedenen Plattformen genutzt werden? Benötigt die Ressource spezifische Systemvoraussetzungen bzw. vorhandene Software (z.B. bestimmte Libraries)?
- Installierbarkeit: Ist die Ressource leicht zu installieren, so dass auch Nutzer ohne Programmierkenntnisse eine Installation durchführen können?
- Ko-Existenz: Kann die Ressource mit anderen Ressourcen gemeinsam genutzt werden oder kann es zu Störungen kommen?

3.2 Diskussion der Anforderungen

Installierbarkeit und **Portabilität** können vernachlässigt werden, wenn die Ressource nur auf Poolrechnern genutzt werden soll und das technische Know-How der Lehrenden oder des lokalen technischen Supports ausreichend ist. Wird erwartet, dass die Studierenden die Ressource auch auf eigenen Rechnern verwenden, werden die beiden Eigenschaften zu entscheidenden Kriterien. Als ideale Lösung bietet sich hier eine **Online-Nutzung** von Ressourcen an. Doch auch diese kann zu Problemen in der Umsetzung führen. Bei WLAN-Nutzung, zum Beispiel, kann eine zu große Gruppe die lokalen Kapazitäten überfordern. Auch auf der Ressourcenseite kann es zu Engpässen kommen, auf die man als externer Nutzer keinen Einfluss hat, wie vorübergehende Ausfälle wegen Wartungsarbeiten. Ein anderes Problem mit reinen Onlineressourcen besteht dann, wenn die Ressource ein sprachverarbeitendes Programm ist und die Studierenden 'private' Sprachdaten, z.B. personenbezogene Texte wie E-Mails oder Blogs, weiterverarbeiten. In diesem Fall ist es aus Datenschutzgründen besser, das verarbeitende Programm ist lokal installiert.

Aus Progammierersicht ist es effizient, wenn ein Programm auf bereits vorhandene Programme und Module zurückzugreift. Dies macht die Installation von Ressourcen jedoch umständlicher und auch schwieriger, weil dann ggf. externe Ressourcen mitinstalliert werden müssen, die eventuell nicht gut dokumentiert sind oder bei denen sich die Verwendung auf verschiedenen Betriebssystemen unterscheidet. Für den Einsatz in der Lehre sind 'Download and Run'-Ressourcen, bei denen die Ressource als nutzungsfähiges **Gesamtpaket** bezogen werden kann, einem flexibel kombinierbaren Modul vorzuziehen.

Die Nutzung eines **Pools** hat den Vorteil, dass die Hardware potenziell homogen ist und die Installation von den Lehrenden durchgeführt werden kann. Klare Nachteile sind die eingeschränkte zeitliche Verfügbarkeit eines Pools und dass es immer Unterschiede zwischen den Privatrechnern der Studierenden und den Poolrechnern geben wird (Betriebssystem, Tastatur, Hilfsprogramme, usw.).

Der Spaßfaktor bei der Nutzung einer Ressource als Ausdruck ihrer **Attraktivität** ist nicht zu unterschätzen. Wann macht es Spaß, mit einer Ressource umzugehen? Eine graphische Oberfläche spielt sicher eine große Rolle. Wenn die Ressource, zum Beispiel, ein Korpus mit Abfrageprogramm ist, dann macht es mehr 'Spaß' damit zu arbeiten, wenn die Suchergebnisse in einer ansprechenden Form optisch dargestellt werden. In TigerSearch kann man, zum Beispiel, einstellen, ob man nur den übereinstimmenden Teilbaum angezeigt haben möchte oder den ganzen Satzgraphen – oder den Satzgraphen mit dem Teilbaum farblich hervorgehoben (in einer Wunschfarbe). Zusätzlich kann der Nutzer wählen, ob und wie viele Kontextsätze angezeigt werden sollen (bis zu drei Kontextsätze).

Indirekt entsteht auch Spaß, wenn der Aufbau einer Ressource oder ihre Handhabung vertraut ist. Der Spaß wird einem vergällt, wenn Erwartungen an einen Inhalt immer wieder enttäuscht werden oder wenn ansonsten automatische Handhabungen ins Leere laufen. Hier spielen manchmal Kleinigkeiten eine große Rolle, zum Beispiel, ob es eine Copy&Paste-Funktionalität gibt, die mit 'normalen' Tastenkombinationen zu bedienen ist. Eine Ressource, die allgemeinen **Konventionen** entspricht, ist leichter zu nutzen als eine sehr individuell gestaltete Ressource, und ist daher in der Lehre vorzuziehen.

Spaß entsteht aber nicht nur, wenn alles bekannt ist und reibungslos abläuft. Im Gegenteil, Spaß entsteht gerade auch dann, wenn ein Prozess des Erkenntnisgewinns stattfindet – wenn man etwas Neues entdeckt oder ein Problem lösen kann, insbesondere dann, wenn am Ende ein 'greifbares' Resultat entsteht.[11]

Für den Einsatz von Sprachressourcen in der Lehre bietet sich als soziale Form die **Gruppenarbeit** an. Probleme können so leichter behoben werden. Vorwissen und Fähigkeiten der Gruppenmitglieder können sich ergänzen. Selbst das gemeinsame Lästern über die Aufgabe oder Probleme bei der Umsetzung tragen zum 'Spaß' bei und können sich indirekt positiv auf das Lernergebnis auswirken.[12]

Ein sehr wichtiges Entscheidungskriterium für die Nutzung einer Ressource in der Lehre ist das Vorhandensein von **Dokumentationen**. Eine gute Dokumentation ist mehrteilig und erfüllt verschiedene Anforderungen: eine allgemeine Beschreibung der Ressource, problem-spezifische Hilfestellung, Beispiele (Beispieldateien bei Programmen oder Suchanfragen und Annotationsbeispiele bei Korpora) und eine detaillierte Nutzungsanleitungen (ein Schritt-für-Schritt-Tutorium oder Annotationsrichtlinien), vgl. auch Dipper et al. (2004).

[11] Dies entspricht dem Ansatz des Forschenden Lernens (vgl. Huber (2004), S. 33, nach Reiber (2007), S. 10).

[12] Das gemeinsame Lästern weckt zumindest passive Teilnehmende auf und löst Emotionen aus, die wiederum grundsätzlich lernfördernd sind. Idealerweise rückt es auch die Handhabung der Ressource ins Zentrum der Aufmerksamkeit.

Die Problem-orientierte Dokumentation kann eine einfache Auflistung im Sinne von häufig gestellten Fragen (FAQs) sein. Bei Korpusressourcen entspricht dies auch der Diskussion von schwierigen Annotationsentscheidungen. Diese Art der Informationspräsentation ist besonders zugänglich für weniger geübte Nutzer, die in einem fortlaufenden Erklärungstext Mühe haben, die Informationen zu finden, die für ihre Fragestellungen relevant sind. Daher ist die **FAQ-Liste** ein wichtiger Bestandteil für denn Einsatz von Sprachressourcen in der Lehre.

Die Bereitstellung von konkreten **Anwendungsbeispielen** erspart den Lehrenden viel Vorbereitungszeit. Dies schließt eine genaue Beschreibung des Formats der Eingabe- und Ausgabedateien von Programmen ein aber auch des Datenformats von Korpusressourcen. Im Idealfall entsteht eine Sammlung von Unterrichtsentwürfen und Sammlungen von Übungsaufgaben, in denen die Ressourcen zum Einsatz kommen. Neben der eigentlichen Dokumentation sind vollständige Beispieldateien sehr hilfreich, mit denen ein Programm unmittelbar getestet werden kann. Im Unterricht können sie als Editiergrundlage für die weitere Arbeit der Studierenden dienen. Die Beispielanwendung dokumentiert idealerweise vollständige Befehle, beim Einsatz von GUIs auch in bildlicher Form (durch Screenshots oder Videotutorien).

Für den Einsatz in der Lehre ist es ebenfalls sehr hilfreich, wenn die Dokumentation auf **Veröffentlichungen** zur Ressource verweist, ebenso auf solche Arbeiten, bei denen die Ressource in irgendeiner Form zum Einsatz kam. In jedem Fall sollte eine einschlägige Referenz angegeben sein, mit welcher die Ressource zitiert werden kann. Im Unterrichtszenario ist dies vor allem dann relevant, wenn das Programm von den Studierenden selbstständig genutzt werden soll z.B. in Abschlussarbeiten.

Die Dokumentation von annotierten Korpora oder lexikalischen Ressourcen beschreibt zusätzlich zu einer möglichen Nutzungsbeschreibung den Inhalt und ggfs. den Entstehungsprozess der Ressource. Bei einem annotierten Korpus kann dies durch die Veröffentlichung der **Annotationsrichtlinien** ('Guidelines') geschehen. Annotationsrichtlinien beinhalten Definitionen der analysierten linguistischen Phänomene, Klassifikationen der Phänomene in abgrenzbare Untertypen und eine exhaustive, eindeutige Zuordnung von Klassen zu Annotationsetiketten (Labeln bzw. 'Tags').

Ein einfaches Annotationsbeispiel sind die verschiedenen Lesarten von *es* (adaptiert von Boyd et al. (2005) und Naumann (2006)). Jede Lesart bildet eine Klasse und wird mit einem spezifischen Etikett versehen:

1. *Nominale Anapher:*
   Das Baby liegt in der Wiege. Es schläft ruhig.
2. *Abstrakte Anapher:*
   Die Benzinpreise steigen wieder. Es ist unglaublich.
3. *Korrelat:*
   Es ist gut, dass Peter kommen konnte.
4. *Wetterverb / Prädikativ der Zeit, des Orts, etc.:*
   ... weil es regnete / ... weil es schon drei Uhr war.
5. *Vorfeld-Es:*
   Es wurde bis zum Morgen getanzt.

Die Annotationsrichtlinien legen auch fest, welche sprachlichen Einheiten überhaupt mit Etiketten markiert werden dürfen ('Markables'). Im Fall der Annotation von *es* als nominale Anapher könnte festgelegt werden, dass nur Nominalphrasen als markierbare Einheiten für die Antezedenten gelten (aber keine Sätze oder ähnliches).

Ein wichtiger Bestandteil von Annotationsrichtlinien sind die Umsetzungskriterien (die Operationalisierungen), wann ein bestimmtes Markable mit einer bestimmten Etikette versehen werden darf. Je konkreter diese Operationalisierungen sind, desto besser können die Nutzer die vorliegenden Annotationen der Daten nachvollziehen. Eine robuste Form der Umsetzung besteht darin, linguistische Tests für das Phänomen zu spezifizieren, z.B. wird im Entscheidungsbaum in Abb. 1 ein Paraphrasierungstest mit *nämlich* vorgeschlagen, um das Antezedenz einer Anapher zu identifizieren[13].

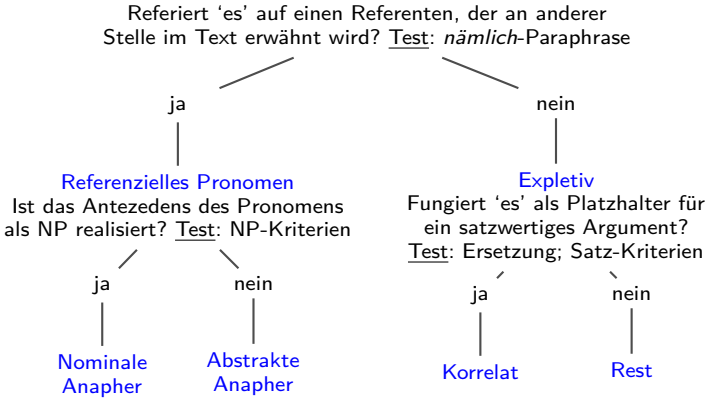


Abbildung 1: Annotationsrichtlinien: Entscheidungsbaum linguistischer Tests.

Neben Definitionen und Tests enthalten die Richtlinien konkrete Annotationsbeispiele: unkontroverse Fälle, die die Analyse veranschaulichen, und Problemfälle, anhand derer die Grenzen der Analyse klar gemacht werden können. Die Richtlinien sollten auch die Annotationsentscheidungen der Problemfälle dokumentieren, so dass die Nutzer nachvollziehen können, welche konkreten Analysen sie in den Daten erwarten.[14]

Sowohl bei Programmen, mit denen bestehende Ressourcen ausgewertet werden, als auch bei solchen, die unmittelbar Daten manipulieren, sollte eine Anleitung vorhanden sein, die angibt, wie das **Analyseergebnis** gesichtet und ggfs. weiterverarbeitet werden kann. Zum Beispiel, wie bestimmte automatische Analyseergebnisse anhand eines Goldstandards evaluiert werden können oder wie die manuelle Annotation von mehreren

[13] Der Paraphrasierungstest zu Beispiel (1): *Es, nämlich das Baby, schläft ruhig.*

[14] Wie man den Annotationsvorgang als didaktisches Mittel in der Lehre einsetzen kann, wird in Zinsmeister (2011) diskutiert.

Annotatoren verglichen werden kann. Bei Korpusressourcen und anderen Daten benötigt man Angaben zu Programmen für die Suche und Visualisierung. Manche Ressourcen stellen APIs für die weitere Prozessierung der (Ausgabe-)Daten zur Verfügung z.B. für eine Konversion in andere Dateiformate. Hier gelten bereits genannte Argumente: Bei Ressourcennutzung auf Privatrechnern sind viele Studierende mit der Verwendung von APIs überfordert. Auch bei einer Poolnutzung ist ein Gesamtpaket leichter zu handhaben.

Bei der Entscheidung, eine bestimmte Ressource in der eigenen Lehre einzusetzen, spielt es eine Rolle, ob man die Ressource auch in der eigenen Forschung verwendet. Es fällt ebenfalls leichter, Ressourcen zu nutzen, die in anderen lokalen Projekten entwickelt oder verwendet werden. Die Wahl zwischen zwei alternativen Ressourcen kann im Zweifelsfall entscheiden, ob man die Entwickler einer Ressource persönlich kennt. Dies alles deutet darauf hin, dass Lehrende Zugriff auf persönliches **Expertenwissen** zu einer Ressource suchen, konkrete Personen, die direkt ansprechbar sind.[15]

In der Lehre ist es ganz entscheidend, die **Motivation**, warum eine bestimmte Ressource genutzt werden soll, zu kommunizieren – was leider immer wieder vernachlässigt wird. Den Studierenden soll das Lernziel bewusst sein und welche Rolle die Ressource in Bezug darauf spielt. Auch wenn das unmittelbare Lernziel eine reine Methoden- oder Ressourcenkompetenz ist, sollten die Studierenden wissen, für welche weiterführenden Forschungsfragen oder Anwendungen die Kompetenz relevant ist. Darüber hinaus sollte man klarstellen, in wie weit die Ressource mittelfristig für das Studium relevant ist, zum Beispiel in anderen Seminaren oder für die studentische Abschlussarbeit. Schlussendlich sollte auch die langfristige Relevanz überprüft werden. Ist das Wissen um die Ressource auch auf Gebiete außerhalb des Studiums übertragbar? Hat es Berufsrelevanz – innerhalb und außerhalb der Forschung?

4 Wunschliste

Aus der Diskussion der Anforderungen oben leiten sich eine Reihe von Wünschen ab, wie man die Bereitstellung von Sprachressourcen für einen Einsatz in der Lehre optimieren könnte. Die Wünsche betreffen nicht den Aufbau der Ressourcen (außer beim Unterpunkt ‘Portabilität‘), sondern ihre Nutzbarmachung für die Lehre.

Bereitstellung Es gibt eine zentrale Webseite mit Links auf Sprachressourcen, die auch als zentrale Informationsstelle dient. Sie bietet Anleitungen im Sinne von Dokumentationen und Expertenrat (eine Art ‘Hotline’). Neben den Ressourcen selbst sammelt sie Forschungsfragen und zeigt Anwendungsbeispiele auf.

[15] In der Praxis wenden sich Nutzer eher an lokale IT-Spezialisten, die selbst oftmals keine Computerlinguisten sind, als an die Entwickler der Ressourcen selbst, vgl. die Diskussion beim D-Spin-Workshop zu eHumanities und Sprachressourcen am 17.01.2011 an der Berlin-Brandenburgischen Akademie in Berlin.

Verfügbarkeit Lehrrelevante Sprachressourcen sind online-nutzbar und stehen auch für einen Download zur Verfügung. Lizenzen werden als Gruppenlizenz für einen ganzen Lehrstuhl vergeben. Die Ressourcen sind für die Lehre kostenfrei. Ein Expertennetzwerk steht für externen Support zur Verfügung.

Verwendbarkeit Die Ressource ist gut dokumentiert. Es gibt eine ausführliche Beschreibung, problem-spezifische Hilfestellung im Sinne von FAQs, Beispieldateien bzw. Annotationsbeispielen sowie eine detaillierte Nutzungsanleitung. Weitere Meta-Dokumentationen benennen Veröffentlichungen zur Ressource sowie Veröffentlichungen, zu der die Ressource beigetragen hat. Zusammen mit der Ressource werden weiterführende Programme zur Verfügung gestellt (siehe Meta-Ressourcen unten).

Portabilität Die Ressource ist leicht zu installieren. Sie wird als plattformunabhängiges Gesamtpaket angeboten.

Dokumentation Wie die Ressourcen selbst sind die Dokumentationen sowohl online zugänglich und damit von den Studierenden immer abrufbar, als auch in der Form eines Dokuments herunterladbar (z.B. als ein pdf-Dokument mit aktiven Links), so dass die Studierenden sie auch lokal speichern und offline nutzen können.

Meta-Ressourcen Bei Programmen werden Konvertierungsprogramme für das geforderte Eingabeformat bereitgestellt, zum Beispiel eine einfache Umformatierung in ein Ein-Wort-Pro-Zeile-Format wie beim TreeTagger. Ebenso wird eine Konvertierung der Enkodierung ermöglicht, zum Beispiel, wenn ein Programm nicht Unicode-basiert arbeitet. Bei Korpora und anderen Daten wird ein Programm zur Recherche und zur Visualisierung der Annotationen angeboten. Unterrichtsbezogene Meta-Ressoucen ergänzen die Bereitstellung: Forschungsfragen, in denen die Ressource eine Rolle spielt, Entwürfe von Lehreinheiten und Übungsaufgaben mit der Ressource. Allgemein – ggf. ganz unabhängig von einzelnen Programmen – kann man in einem Webformular die Bewertung von Annotationen gegen einen beliebigen Goldstandard evaluieren bzw. die Inter-Annotatoren-Übereinstimmung berechnen, sowie sich Konfusionsmatrizen anzeigen lassen, die darstellen, welche Annotationsetiketten wie oft miteinander verwechselt wurden. Ebenso kann man über ein allgemeines Webformular Kodierungskonversionen online durchführen (z.B. utf-8 nach iso-latin-1).

An wen richten sich die genannten Wünsche? Die Entwickler der Ressourcen sind damit sicher überfordert. Die Umsetzung sollte vielmehr an zentralen Stellen geschehen, an denen Ressourcen und lehrrelevante Meta-Ressourcen nachhaltig bereitgestellt werden können.

5 Zusammenfassung und Ausblick

Mehrfach war angeklungen, dass die Online-Nutzung von Ressourcen eine gewünschte Option für die Lehre sei. Diese Idee war früh in der *Linksammlung zu computerlinguistische Online-Ressourcen* an der Universität Zürich umgesetzt worden.[16] Der Online-Service *WebLicht*[17] des Projekts D-Spin geht noch einen Schritt weiter. Es bietet ein online nutzbares Portal zur Verarbeitung von Sprachdaten an. Nutzer laden ihre Daten in das System und stellen sich per Drag&Drop eine Verarbeitungsabfolge aus verschiedenen Programmen zusammen, zum Beispiel aus einem Textsegmentierer, einem Wortartentagger und einem syntaktischen Parser. Der angereicherte Text wird anschließend wieder auf den lokalen Rechner gespeichert. Ressourcenentwickler sind aufgerufen, als Partner ihre Programme in Weblicht einzuspeisen. Eine andere Art von Online-Sammlung stellt die *Studienbibliographie Computerlinguistik* dar.[18] Thematisch strukturiert nennt sie vorwiegend Literatur, weist aber auch auf relevante Sprachressourcen hin. Das *Portal Computerlinguistik*[19], ein Gemeinschaftsprojekt der Gesellschaft für Sprachtechnologie und Computerlinguistik (GSCL) und der Computerlinguistik-Sektion der Deutschen Gesellschaft für Sprachwissenschaft (DGFS-CL), ist noch im Aufbau. Es hat das Ziel, Informationen und Ressourcen zur Sprachverarbeitung mit einem Schwerpunkt auf der Verarbeitung des Deutschen bereitzustellen. Einen speziellen Support für die Nutzung der Ressourcen in der Lehre gibt es bisher nicht.

Über den deutschsprachigen Tellerrand hinaus ist die *Open Language Archives Community* (OLAC) zu nennen, die eine weltweite virtuelle Bibliothek von Sprachressourcen aufbaut[20] sowie die Web-Links der Linguist List[21].

Müssen Lehrende einen computerlinguistischen Hintergrund besitzen, um Sprachressourcen in der Lehre einsetzen zu können? Auch in rein sprachwissenschaftlichen Seminaren können Studierende von Sprachressourcen wie Online-Korpora oder Online-Parsern profitieren, zu deren Nutzung man keine eigentliche computationelle Kompetenz benötigt. Die Ressourcen können helfen – wie von der Lernpsychologie gefordert – Sprachdaten und Analysen 'greifbar' zu machen.

Danksagung

Vielen Dank an die anonymen Gutachter für Korrekturen und hilfreiche Kommentare. Diese Arbeit ist gefördert aus dem Europäischen Sozialfonds in Baden-Württemberg.

[16] Züricher Sammlung zu Online-Demos: `http://kitt.cl.uzh.ch/kitt/cltools`

[17] WebLicht (`https://weblicht.sfs.uni-tuebingen.de/`) ist offen für Mitglieder der 'Clarin Identity Federation'. Viele Universitäten und andere Hochschule besitzen diesen Status.

[18] Studienbibliographie CL: `http://www.coli.uni-saarland.de/projects/stud-bib/`.

[19] Portal CL: `http://www.computerlinguistik.org/portal/portal.html?s=Home`.

[20] OLAC: `http://www.language-archives.org/`.

[21] Linguist List: `http://linguistlist.org`, zum Beispiel unter 'Education' die Links 'Software' und 'Linguistic Exercises and Aids'.

Literatur

Aston, G. (2000). Learning English with the British National Corpus. In Battaner, M. and López, C., editors, *VI jornada de corpus lingüístics*, pages 25–40. Barcelona.

Boyd, A., Gegg-Harrison, W., and Byron, D. (2005). Identifying non-referential it: A machine learning approach incorporating linguistically motivated patterns. In *Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in Natural Language Processing*, pages 40–47, Ann Arbor, Michigan. Association for Computational Linguistics.

Dipper, S., Götze, M., and Stede, M. (2004). Simple annotation tools for complex annotation tasks: an evaluation. In *Proceedings of the LREC Workshop on XML-based Richly Annotated Corpora*, pages 54–62, Lisbon.

EAGLES (1996). Evaluation of natural language processing systems. Final report. EAGLES DOCUMENT EAG-EWG-PR.2. http://issco-www.unige.ch/projects/ewg96/ewg96.html.

Huber, L. (2004). Forschendes Lernen: 10 Thesen zum Verhältnis von Forschung und Lehre aus der Perspektive des Studiums. *Die Hochschule*, 2:29–49.

Müller, C. and Strube, M. (2006). Multi-level annotation of linguistic data with MMAX2. In Braun, S., Kohn, K., and Mukherjee, J., editors, *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, pages 197–214. Peter Lang, Frankfurt a.M.

Naumann, K. (2006). *Manual of the Annotation of in-document Referential Relations*. University of Tübingen.

Poesio, M. (2000). Coreference. In Mengel, A., Dybkjaer, L., Garrido, J., Heid, U., Klein, M., Pirrelli, V., Poesio, M., Quazza, S., Schiffrin, A., and Soria, C., editors, *MATE Deliverable D2.1. MATE Dialogue Annotation Guidelines*, pages 134–187.

Rafferty, A. and Manning, C. D. (2008). Parsing three German treebanks: Lexicalized and unlexicalized baselines. In *Proceedings of the Workshop on Parsing German*, pages 40–46, Columbus, Ohio. Association for Computational Linguistics.

Reiber, K. (2007). Grundlegung: Forschendes Lernen als Leitprinzip zeitgemäßer Hochschulbildung. *Tübinger Beiträge zur Hoschuldidaktik: Forschendes Lernen als hochschuldidaktisches Prinzip – Grundlegung und Beispiele*, 3(1):6–12.

Schiller, A., Teufel, S., Stöckert, C., and Thielen, C. (1999). Guidelines für das Tagging deutscher Textcorpora mit STTS. Technical report, Institut für maschinelle Sprachverarbeitung, Stuttgart.

Schmid, H. (1997). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *New Methods in Language Processing*, pages 154–164. UCL Press, London.

Schmidt, T. (2004). Transcribing and annotating spoken language with EXMARaLDA. In *Proceedings of the LREC-Workshop on XML based richly annotated corpora, Lisbon 2004*, Paris. ELRA.

Winteler, A. (2004). *Professionell lehren und lernen. Ein Praxisbuch*. Wissenschaftliche Buchgesellschaft.

Zinsmeister, H. (2011). Exploiting the 'annotation cycle' for teaching linguistics. Vortrag beim Workshop Corpora in Teaching Languages and Linguistics. Berlin. http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/events-en/CTLL/ctll_abstracts/ctll_zinsmeister.

Stefanie Dipper

Digitale Korpora in der Lehre — Anwendungsbeispiele aus der Theoretischen Linguistik und der Computerlinguistik

In diesem Artikel werden verschiedene Szenarien aus der Lehre vorgestellt, in denen Korpora (und andere Sprachressourcen) Einsatz finden. Jeweils zwei Beispiele illustrieren die Nutzung von Korpora in der Theoretischen Linguistik und in der Computerlinguistik. In der Theoretischen Linguistik dienen Korpora als Belegquellen oder Testdaten für die Hypothesen aus der theoretischen Forschung. In der Computerlinguistik werden Korpora für die Anwendungsentwicklung oder für den Ressourcenaufbau eingesetzt.

1 Einführung[1]

In diesem Artikel geht es um den Einsatz digitaler Sprachressourcen in der universitären Lehre. Die hier angesprochenen Sprachressourcen sind unterschiedlicher Art und reichen von "rohen Texten", wie man sie z.B. im Internet findet, über sorgfältig aufbereitete, annotierte Korpora und spezialisierte Korpus-Suchtools bis hin zu automatischen Analysetools, z.B. für die automatische Wortart-Erkennung. Die Lehrveranstaltungen, in denen diese Sprachressourcen zum Einsatz kommen, richten sich zum einen an Studenten der Theoretischen Linguistik, zum andern an Studenten der Computerlinguistik.

Es werden insgesamt vier Erfahrungsberichte aus der Lehre vorgestellt. Sie sind so ausgewählt, dass sie möglichst verschiedenartige Aspekte von Sprachressourcen, ihrer Nutzung und der an sie gestellten Anforderungen illustrieren. Konkret werden zwei Einsatzszenarien aus der Theoretischen Linguistik (Abschnitt 2) und zwei aus der Computerlinguistik (Abschnitt 3) beschrieben. In Abschnitt 4 folgen abschließende Anmerkungen.

2 Korpora in der Theoretischen Linguistik

In den ersten beiden Erfahrungsberichten geht es um (kleinere) Lehreinheiten, wie sie im Rahmen eines Bachelor-Studiengangs für Studenten der Linguistik angeboten werden.

2.1 Korpora als Belegquellen

Im ersten vorgestellten Einsatzszenario dienen Korpora vorwiegend als Belegquellen. In den Korpora wird nach sprachlichen Belegen gesucht, die Instanzen eines im Kurs behandelten linguistischen Phänomens sind.

[1] Alle angegebenen URLs wurden am 8.3.2011 abgerufen.

Als ein Beispiel für ein solches Phänomen soll die *Vorfeldbesetzung* im Deutschen dienen. Dabei geht es um die Frage, was für Faktoren bestimmen, welche Konstituente in einem deutschen Hauptsatz die Position vor dem finiten Verb, das sogenannte *Vorfeld*, einnimmt. Im Gegensatz zum Englischen kann hier eine beliebige maximale Phrase stehen, z.B. eine Subjekt- oder Objekt-NP, eine PP oder auch eine Adverbialphrase, wie in Beispiel (1)[2] (die Vorfeld-Konstituente ist unterstrichen).

(1) Manchmal mußten erst Mahnschreiben in Sahlins Briefkasten landen, bevor die mit rund 10.000 Mark monatlich nicht unbedingt schlecht versorgte Ministerin sich zur Rückzahlung bequemte.

Die Faktoren, die hier eine Rolle spielen, sind recht gut untersucht: Neben der grammatischen Funktion der Vorfeld-Konstituente (es sind mehrheitlich Subjekte) spielt auch (Nicht-)Vorerwähntheit eine wichtige Rolle (z.B. Filippova and Strube (2007); Speyer (2007)). So wird die Vorfeldposition u.a. auch bevorzugt von Ausdrücken eingenommen, die für den Hörer neue Information darstellen, und von sogenannten *scene-setting elements* (zu denen vermutlich auch Bsp. (1) gerechnet werden kann).

Es liegt nun nahe, die Studenten — nach einer Einführung in das Thema und die involvierten Faktoren — gezielt nach Belegen suchen zu lassen, die gegen die genannten "Regeln" verstoßen, also z.B. nach Sätzen, in denen ein vorerwähntes Objekt im Vorfeld steht. Solche Beispiele, die potenziell die linguistischen Hypothesen widerlegen, können dann gezielt untersucht und die Hypothesen gegebenenfalls verfeinert werden.

Für eine effiziente Suche nach Belegen sind zwei Dinge notwendig: (i) entsprechend annotierte Korpora und (ii) Korpus-Suchtools, mit denen diese Annotationen abgefragt werden können.

Korpora Für unsere Beispielsuche von oben — "vorerwähntes Objekt im Vorfeld" — benötigen wir eine *Baumbank*, d.h. ein Korpus, das mit syntaktischer Information (Konstituenten mit ihren Kategorien und Funktionen) annotiert ist. Zusätzlich wollen wir die Position im Vorfeld sowie Vorerwähntheit abfragen können. Tabelle 1 listet die aktuell verfügbaren Baumbanken für das Deutsche[3] und gibt an, ob die Zusatzmerkmale im Korpus mit annotiert sind (die eingeklammerten Werte werden unten näher erläutert).

Korpus-Suchtools Die Korpora liegen typischerweise in mehreren Formaten vor, z.B. im NEGRA-Exportformat (einem Datenbank-ähnlichen Spaltenformat, Brants (1997))

[2] Entnommen aus der TüBa-D/Z: `http://www.sfs.uni-tuebingen.de/tuebadz.shtml`

[3] TüBa-D/Z: "Tübinger Baumbank des Deutschen/Zeitungskorpus", `http://www.sfs.uni-tuebingen.de/tuebadz.shtml`
TüBa-D/S: "Tübinger Baumbank des Deutschen/Spontansprache" (Projekt Verbmobil), `http://www.sfs.uni-tuebingen.de/tuebads.shtml`
NEGRA-Korpus: Projekt "Nebenläufige grammatische Verarbeitung" (SFB 378), `http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/`
TIGER-Korpus: "Linguistic Interpretation of a German Corpus", `http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERCorpus/`
PCC: "Potsdam Commentary Corpus", `http://www.ling.uni-potsdam.de/pcc/pcc.html`

Baumbank	Vorfeld	Vorerwähntheit	#Sätze	#Tokens
TüBa-D/Z	✓	✓	55.000	980.000
TüBa-D/S	✓	–	38.000	360.000
NEGRA-Korpus	(✓)	–	20.000	350.000
TIGER-Korpus	(✓)	–	50.000	900.000
PCC	(✓)	✓	2.900	44.000

Tabelle 1: Deutsche Baumbanken und die Merkmale 'Vorfeld' und 'Vorerwähntheit'. Der Eintrag '(✓)' bedeutet, dass das Merkmal nicht explizit annotiert ist, aber "simuliert" werden kann. In den letzten Spalten finden sich Angaben zur Größe der Korpora.

oder in TIGER-XML (Mengel and Lezius, 2000). Um Linguisten einen einfachen Zugang zu den Daten zu ermöglichen, wurden spezialisierte Korpus-Suchtools entwickelt. Diese erlauben es dem Nutzer, Anfragen an das Korpus in einer Anfrage-Sprache zu formulieren, die auf Grundkonzepten der Linguistik beruht. Beispielsweise stellen solche Anfrage-Sprachen Operatoren für die Relationen *Dominanz* und *Präzedenz* bereit. Neben der Funktionalität der *Suche* auf den Daten bieten Korpustools außerdem geeignete *Visualisierungen* der Daten an, z.B. in Form von Phrasenstruktur-Bäumen.

Ein oft verwendetes Suchtool für Baumbanken ist TIGERSearch[4], das unter vielen anderen Formaten auch Daten in TIGER-XML verarbeiten kann. Dieses Tool hat allerdings den (in unserem Fall relevanten) Nachteil, dass es satzweise operiert. D.h. mit TIGERSearch kann nur nach Relationen innerhalb eines Satzes gesucht werden, und das Tool kann nur Bäume für einzelne Sätze darstellen.

Ein Suchtool, das satzübergreifende Abfragen erlaubt, ist ANNIS2[5], das als Multifunktionstool neben baumartigen Strukturen auch Annotationen von Spannen und Zeigerstrukturen erfasst. Die Baumbank TüBa-D/Z, in der das Merkmal der Vorerwähntheit (in Form von Anaphern- und Koreferenz-Relationen) annotiert ist, kann mit ANNIS2 abgefragt werden.

Die in (2) gezeigte Anfrage sucht nach vorerwähnten Objekten im Vorfeld: Zuerst wird allgemein nach Knoten der Kategorie 'VF' (= Vorfeld, Zeile 1) und nach Knoten der Kategorie 'NX' (= Nominalphrasen/NPs, Zeile 2) gesucht. Als zusätzliche Einschränkung wird angegeben, dass die NP als 'OA' (= Objekt im Akkusativ) fungiert und im Vorfeld steht (Zeile 4); die Angaben '#1' und '#2' beziehen sich dabei auf die Knoten, die in Zeile 1 bzw. 2 spezifiziert wurden.[6] Schließlich, als weitere Einschränkung, wird eine

[4] `http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/`

[5] ANNIS: "Annotation of Information Structure, `www.sfb632.uni-potsdam.de/annis/`

[6] Der Ausdruck sucht nach einem Knoten der Kategorie 'VF', der einen Knoten der Kategorie 'NX' dominiert, wobei die Dominanz-Kante ebenfalls ein Label erhält: der 'NX'-Knoten fungiert als 'OA'. Dass die Position 'Vorfeld' als Kategorie-Label annotiert wird, kommt für linguistische Nutzer

nicht näher spezifizierte 'relation' (= Anaphern- oder Koreferenz-Relation) eingeführt (Zeile 3), die auf die NP zutreffen soll (Zeile 5); mit anderen Worten: die NP soll also in irgendeiner Form bereits vorerwähnt sein.

(2)
```
cat="VF" &
cat="NX" &
relation=/.*/ &
#1 >[func="OA"] #2 &
#2 _=_ #3
```

Diese Anfrage findet 139 Treffer im Korpus, darunter den Treffer, der in Abb. 1 gezeigt wird.[7] Die so gefundenen Belege können nun einzeln durchgesehen und genauer untersucht werden.[8]

Für das PCC-Korpus, das ebenfalls eine Koreferenz-Annotation enthält, sieht die entsprechende Anfrage in ANNIS2 wie in (3) aus. Die Form des Suchausdruckes unterscheidet sich deutlich von dem in (2). Dafür gibt es mehrere Gründe: Das PCC verwendet das Syntax-Annotationsschema des TIGER-Korpus[9], in dem keine explizite Markierung des Vorfeldes vorgesehen ist. Stattdessen machen wir die (schwächere) Einschränkung, dass das Objekt die *linke Tochter* des Satzes oder der VP ist (Zeile 5: '@l'). Außerdem gibt es keine einheitliche Auszeichnung für NPs in TIGER; daher lassen wir die Kategorie unterspezifiziert ('node', Zeilen 2–4). Schließlich unterscheidet sich die Form der Koreferenz-Annotation in beiden Korpora (Zeile 6).

(3)
```
cat=/(S|VP)/ &
node &
node &
node &
#1 >@l[func="OA"] #2 &
#3 ->anaphor_antecedent #4 &
#2 _=_ #3
```

sicher zunächst unerwartet. Solche Annotationsentscheidungen müssen den Annotationsguidelines entnommen werden, die als Dokumentation ein unverzichtbarer Bestandteil des Korpus sind (für die TüBa-D/Z: `http://www.sfs.uni-tuebingen.de/resources/tuebadz-stylebook-0911.pdf`).

[7] Die Abfrage in (2) wurde nicht auf der aktuellen Version 6, sondern auf der Vorgängerversion 5 der TüBa-D/Z durchgeführt, die 800.000 Tokens enthält.

[8] Allerdings muss bei solchen Untersuchungen berücksichtigt werden, dass das Konzept von Vorerwähntheit, wie es in der TüBa-D/Z annotiert ist, nicht notwendigerweise mit dem Konzept, wie es in den linguistischen Theorien verstanden wird, identisch ist. Z.B. ist die Vorfeld-NP *das Geld* in Bsp. (i) laut den Annotationen der TüBa-D/Z nicht "vorerwähnt" (da das Geld *als solches* noch nicht thematisiert wurde). Im Vorkontext ist jedoch schon die Rede von "bezahlen", daher stellt der Ausdruck *das Geld* natürlich keine völlig neue (unerwartete) Information im Sinne von Filippova and Strube (2007) oder Speyer (2007) dar.

(i) Seit 1991 bezahlte sie mal neue Schuhe, eine Lederjacke oder gleich die Urlaubsreise für die ganze Familie mit der Staatskarte. Das Geld zahlte sie zurück, aber dummerweise nicht gleich und nicht unaufgefordert.

[9] `http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERCorpus/annotation/tiger_scheme-syntax.pdf`

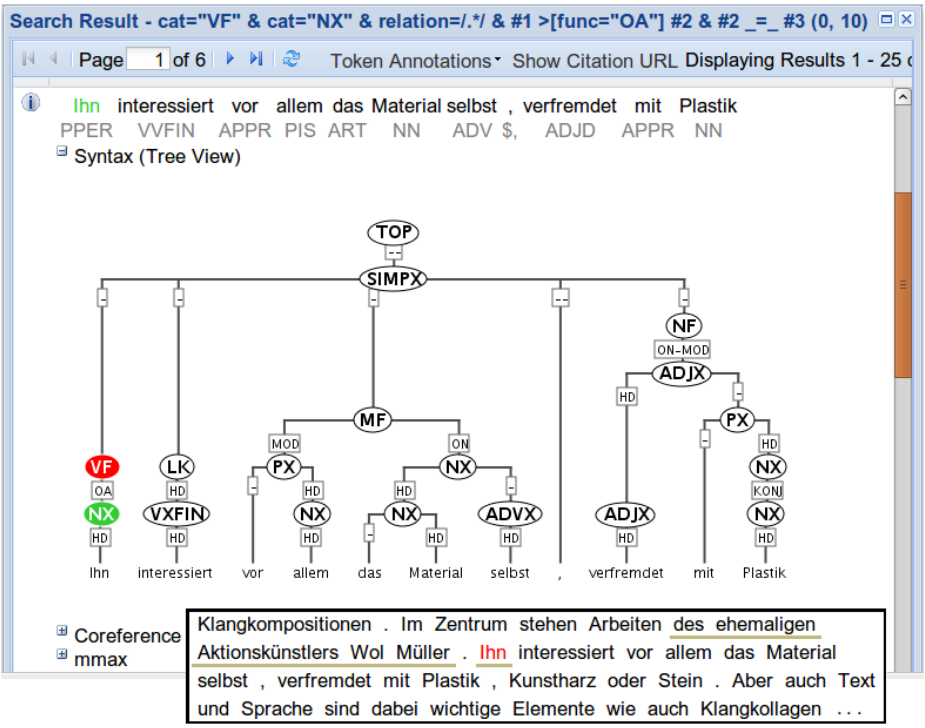


Abbildung 1: Screenshot von ANNIS2: Treffer der Suche nach einem vorerwähnten Objekt im Vorfeld (aus TüBa-D/Z). ANNIS2 markiert die Knoten 'VF' und 'NX', die in der Anfrage spezifiziert wurden, im Baum farbig. In der eingefügten Box unten ist ein Ausschnitt aus der Koreferenzansicht von ANNIS2 gezeigt, koreferente Ausdrücke werden hier mit der gleichen Farbe unterstrichen (im Beispiel: *des ehemaligen Aktionskünstlers Wol Müller* und *ihn*).

Zugang zu den Ressourcen In dem oben geschilderten Szenario sollen Studenten selbständig in Korpora nach relevanten Belegen suchen. Voraussetzung dafür ist zum einen, dass die Korpora und Suchtools frei verfügbar sind (für die Lehre). Alle oben aufgeführten Korpora und Tools erfüllen diese Bedingung: sie sind für nicht-kommerzielle Anwendungen kostenlos lizenzierbar.

Zum andern sollten die technischen Hürden möglichst niedrig sein, sowohl bei der Installation wie auch bei der Bedienung der Tools. Zum aktuellen Zeitpunkt bieten sich dafür Tools an, die auf einem Server installiert und von den Studenten über einen Webbrowser bedient werden können. Auf diese Weise müssen Prozesse wie das Installieren des Suchtools oder das Aufbereiten und Einlesen der Korpora nicht unzählige Male durchgeführt werden (nämlich von jedem einzelnen Studenten auf dem eigenen Rechner), sondern passieren einmalig auf dem zentralen Server.

Von den oben erwähnten Korpustools ist ANNIS2 ein solches webbasiertes Tool. Ein weiteres Beispiel ist CQPweb[10], das allerdings etwas andere Funktionalitäten als ANNIS2 bietet: Es unterstützt keine Baum- oder Zeigerstrukturen, sondern nur Wort- oder Spannen-basierte Annotationen; andererseits können in CQPweb auch Parallelkorpora abgefragt werden.

Eine andere Möglichkeit ist, Korpora zu nutzen, die zwar nicht lizenzierbar sind, aber über ein Web-Interface abgefragt werden können. Dazu zählen z.B. die DWDS-Korpora, die mit POS und Lemmata annotiert sind, oder COSMAS II, dessen Korpora mit POS, Lemmata und Morphologie annotiert sind.[11]

Approximationen Bei Korpora, die z.B. nur mit Wortarten ('parts of speech', POS) annotiert in CQPweb vorliegen, lassen sich syntaktische Anfragen mit Hilfe *regulärer Ausdrücke* über den POS-Annotationen approximieren. Der Suchausdruck in (4), formuliert in der Anfrage-Sprache CQP, erfasst beispielsweise eine Teilmenge der Akkusativ-NPs am Satzanfang, denen ein finites Verb folgt.[12]

(4) `[pos="ART" & word="(Den|Einen)"][pos="ADJA"]*[pos="N."][pos="V.FIN"]`

Schon in der Anfrage in (3) haben wir diese "Technik" genutzt, nicht explizit annotierte Information mit Hilfe der vorhandenen Annotation zu approximieren. Die Annotationen im NEGRA-, TIGER- und dem PCC-Korpus erlauben eine gute Approximation der Vorfeld-Position, daher ist dieses Merkmal in Tabelle 1 mit '(✓)' markiert.

[10] Web-Interface für den "Corpus Query Processor", `http://cwb.sourceforge.net/`. Über CQPweb wird auch eine Reihe von Demo-Korpora angeboten, `http://cwb.sourceforge.net/demos.php`.

[11] DWDS: "Digitales Wörterbuch der Deutschen Sprache", `http://www.dwds.de/`
COSMAS: "Corpus Search, Management and Analysis System", `https://cosmas2.ids-mannheim.de/cosmas2-web/`

[12] Die POS-Tags in (4) folgen dem STTS (Schiller et al., 1999): 'ART' markiert Artikel, 'ADJA' attributive Adjektive, 'N.' steht für 'NN' oder 'NE', d.h. für allgemeine Nomen, 'V.FIN' für 'VAFIN', 'VMFIN' oder 'VVFIN', d.h. für finite Verben. Die abgefragte NP beginnt mit einem definiten oder indefiniten maskulinen Artikel, da dieser eindeutig für Akkusativ markiert ist; die Großschreibung bewirkt, dass nur Artikel am Satzanfang gefunden werden. Danach stehen beliebig viele Adjektive, gefolgt vom Kopfnomen. Der Ausdruck erfasst also beispielsweise weder feminine noch artikellose NPs oder NPs mit postnominalen Modifikatoren.

Die Beispiele in diesem Abschnitt illustrieren, dass das linguistische Wissen, das bei der Suche nach relevanten Belegen nötig ist, auf unterschiedliche Weise einfließen kann: Entweder das linguistische Wissen ist bereits im Korpus vorhanden, in Form von komplexen Annotationen, die nur noch abgefragt werden müssen. Oder das Korpus enthält nur einfache Annotationen, die in einer komplexen Anfrage miteinander kombiniert werden; hier verteilt sich das linguistische Wissen auf Annotation und Abfrage. Schließlich sichtet man gegebenenfalls die Treffer von Hand, was ebenfalls den Einsatz von linguistischem Wissen erfordert.

2.2 Korpora als "Testdaten"

Im zweiten Einsatzszenario geht es um ein linguistisches Phänomen, das nicht so deutlich abgrenzbar ist wie die Vorfeldbesetzung: die Realisierung des *(Aboutness-)Topiks*. In der linguistischen Forschung wird als ein Test für die Bestimmung des Topiks eines Satzes die Einleitungsphrase *Ich erzähle dir etwas über X* genannt. Die Konstituente, die an Stelle des 'X' eingesetzt werden kann, ohne dass der Sinn des nachfolgenden Satzes verändert wird, ist die Topik-Konstituente des Satzes. Im Beispiel (5) (entnommen aus Frey (2004)) ergibt der Test, dass *Maria* die Topik-Konstituente ist (im Beispiel unterstrichen). Als alternative Einleitungsphrasen werden genannt: *Was ist mit X?* oder *Was X anbelangt, ...* (Götze et al., 2007).

(5) (Ich erzähle dir etwas über Maria.)
Nächstes Jahr wird Maria wahrscheinlich nach London gehen.

Topiks sind oft Subjekte, außerdem stehen sie oft im Vorfeld oder am Anfang des Mittelfeldes (Frey, 2004). Geht man allerdings weg von den konstruierten Beispielsätzen aus der Forschungsliteratur (wie Bsp. (5)) hin zu Alltagstexten, stellt man schnell fest, dass die Bestimmung des Satztopiks oftmals nicht trivial ist. Angewendet auf fortlaufenden Text führen die genannten Tests mit Einleitungsphrasen häufig zu unnatürlichen Ergebnissen und unklaren Intuitionen. (6) zeigt einen Beispielsatz im originalen Kontext, der zwei gleichwertige Topik-Kandidaten enthält (entnommen aus Cook and Bildhauer (2011)).

(6) (Dazugelernt habe ich besonders im Bereich der Öffentlichkeitsarbeit. Ich merkte, welche Handlung welche Reaktion auslöst und wie man gewisse Ereignisse richtig kommuniziert.)
Von dieser Erfahrung$_{top?}$ kann ich$_{top?}$ am neuen Ort selbstverständlich profitieren.

D.h. im Vergleich zum Begriff des Vorfeldes ist der Begriff des Topiks vage und schwierig anzuwenden. Entsprechend gibt es bislang auch nur wenige Versuche, Korpora mit Topiks zu annotieren — und folglich stehen für die Lehre erstmal keine vorannotierten Korpora zur Verfügung. Eine Möglichkeit ist daher, die Studenten nach einer theoretischen Einführung in die Thematik selbst Texte annotieren zu lassen. Für die (manuelle) Annotation benötigt man, neben dem zu annotierenden Text, (i) ein Annotationstool und (ii) Annotationsguidelines.

Annotationstools Der Einsatz von Annotationstools im Kurs hat mehrere Vorteile gegenüber der Annotation auf Papier: Die Studenten lernen den Umgang mit den Tools; die Daten können nachhaltig gespeichert werden (und z.B. in nachfolgenden Sitzungen weiter verwendet werden); Daten können kollaborativ annotiert werden; die Annotatoren-Übereinstimmung kann berechnet und kontrovers annotierte Sätze können automatisch bestimmt werden.

Aus den schon oben genannten Gründen gilt auch hier, dass das Annotationstool möglichst webbasiert operieren sollte, also der Zugang und die Bedienung über einen Webbrowser möglich sein sollte. Da der Nutzer hier eigene Texte annotieren will, muss das Tool eine Upload-Funktion oder eine Texteingabe anbieten. Neben einer Download-Funktion wäre außerdem ein direkter Weg vom Annotationstool in ein Korpus-Suchtool ideal, in dem die annotierten Daten gleich durchsucht werden können.

Für den gelegentlichen Einsatz im Linguistik-Unterricht ist es wichtig, dass das Annotationstool möglichst einfach und intuitiv bedient werden kann. Das Annotations-Interface sollte sich daher an gängigen Editoren orientieren (z.B. in der Verwendung von Tastaturbefehlen). Eine Unterstützung des Text-Uploads durch einen integrierten Tokenisierer ist denkbar.

Zum aktuellen Zeitpunkt ist der überwiegende Teil der verfügbaren Annotationstools als *stand alone*-Anwendung realisiert, die lokal installiert werden muss. Zu den wenigen Ausnahmen gehören die Tools Serengeti, das für die Annotation von Koreferenz-Beziehungen entwickelt wurde, und Typecraft, mit dem wortbasierte Annotationen erstellt werden können.[13] Daneben werden seit einiger Zeit Web-Interfaces für (meist kommerzielle) *crowd sourcing*-Annotationen entwickelt.

Annotationsguidelines Werden die bekannten Tests als Kriterien für die Annotation genommen, so werden die Studenten schnell feststellen, dass die Tests häufig nur unbefriedigende Ergebnisse liefern und die Intuitionen oft unklar bleiben. Eine (anspruchsvolle) Aufgabe für die Studenten könnte dann darin bestehen, das Konzept 'Topik' besser zu operationalisieren, d.h. bessere linguistische Tests dafür zu entwickeln. Angewendet auf fortlaufende Alltagstexte und sukzessive erweitert und verfeinert, können diese Tests schließlich in "robuste" Guidelines münden, die für alle denkbaren Fälle eindeutige Kriterien der Annotation vorsehen.

Auf diese Weise wird das Korpus letztlich zum "Testkorpus" für die linguistische Theorie: Mit einer wachsenden Menge von klar definierten Kriterien und von annotierten Daten ergeben sich Hinweise darauf, welche sprachliche Realität Konzepte wie Topik im Deutschen bzw. in Korpora des Deutschen haben. Ein annotiertes Korpus kann außerdem als Test für die Vorhersagen der Theorie dienen.

Ein kurzes Zwischenfazit: Im ersten Szenario war das Phänomen klar umrissen (Vorfeld-Besetzung); der Fokus lag auf der Suche nach markierten Einzelbelegen in annotierten Korpora. Dagegen ist im zweiten Szenario das Phänomen schwerer fassbar (Topik) und

[13] Serengeti: "Semantic Relations Annotation Tool", `http://coli.lili.uni-bielefeld.de/serengeti/`
Typecraft: `http://www.typecraft.org/`

die linguistische Analyse noch unklar; der Fokus liegt daher auf der Untersuchung und Annotation von fortlaufendem Text.

Für beide Szenarien gilt, dass die Annotations- und Suchtools satzübergreifende Relationen mit abdecken und die Bedienung webbasiert erfolgen sollte. Idealerweise sollte der Output der Annotationstools direkt in ein Suchtool eingespeist werden können.

3 Korpora in der Computerlinguistik

Die beiden nächsten Erfahrungsberichte stammen aus einem Bachelor-Studiengang für Computerlinguistik. Hier geht es um kleinere Implementationsaufgaben, die von den Studenten weitgehend eigenständig bearbeitet werden (im Rahmen sogenannter "Forschungsprojekte").

3.1 Korpora in der Anwendungsentwicklung

Als Anwendungsbeispiel soll die Aufgabe der *Autorenschaft-Zuschreibung* dienen: Gegeben eine Menge an Texten von verschiedenen Autoren wird ein Klassifikator gesucht, der einen neuen Text einem der Autoren zuweist.

Vorgabe im Kurs war es, nicht einen reinen *bag of words*-Ansatz zu implementieren, bei dem der Klassifikator ausschließlich *N-Gramme* (Sequenzen) von Wortformen nutzt. Stattdessen sollten auch linguistische Merkmale wie Wortart (POS), Morphologie o.ä. eine Rolle spielen.

Die erste Aufgabe der Studenten bestand darin, sich nach geeigneten Ressourcen umzusehen. Das betraf zum einen die Textgrundlage, d.h. eine Menge von geeigneten, vergleichbaren Texten, bei denen die Autorenschaft bekannt ist. Zum andern betraf es automatische Analysetools, die zu diesen Texten die linguistischen Merkmale für den Klassifikator liefern sollten.

Im Folgenden werden zwei Studentenprojekte mit verschiedenen Textgrundlagen vorgestellt.

Beispiel-Korpus: Enron Eine Gruppe wählte als Korpus das *Enron Email Dataset* (Klimt and Yang, 2004; Shetty and Adibi, 2004). Enron war ein amerikanischer Energiekonzern, der nach umfangreichen Bilanzfälschungen 2001 Insolvenz anmelden musste. Im Zuge der Ermittlungen gegen Enron stellte die amerikanische Bundesaufsichtsbehörde für Energie den (privaten wie beruflichen) Email-Verkehr von 150 leitenden Enron-Mitarbeitern im Internet frei zur Verfügung. Von verschiedenen Forschungsinstituten wurden diese Daten für Forschungszwecke weiter aufbereitet, indem z.B. Duplikate und leere Mails entfernt wurden. Diese Daten, das Enron Email Dataset, können frei heruntergeladen werden.[14] Das Korpus enthält die Ordner der Posteingänge der Mitarbeiter, d.h. die Emails sind nach ihren Empfängern sortiert. Abb. 2 zeigt eine Email, wie sie im Korpus enthalten ist.

[14] Z.B. unter `http://www.cs.cmu.edu/~enron/`

```
Date: Tue, 5 Dec 2000 02:51:00 -0800 (PST)
From: denise.lagesse@enron.com
To: drew.fossum@enron.com
Subject: Susan's expense report 11/16/00
Cc: susan.scott@enron.com
Mime-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit
Bcc: susan.scott@enron.com
X-From: Denise LaGesse
X-To: Drew Fossum
X-cc: Susan Scott
X-bcc:
X-Folder: \Drew_Fossum_Dec2000_June2001_1\Notes Folders\Notes inbox
X-Origin: FOSSUM-D
X-FileName: dfossum.nsf

Drew,

Did you already approve this expense report and send it to accounting?

Denise

---------------------- Forwarded by Denise LaGesse/ET&S/Enron on 12/05/2000
10:50 AM ---------------------------


DENISE LAGESSE
11/16/2000 04:04 PM
To: Drew Fossum/ET&S/Enron@ENRON
cc: Susan Scott/ET&S/Enron@ENRON

Subject: Susan's expense report 11/16/00

Please approve the attached expense report and cc: me on your approval.
Thanks!
```

Abbildung 2: Email aus dem Enron-Korpus. Es handelt sich um die Datei 'fossum-d/notes_inbox/103' aus dem Release vom 21. August 2009. Im Email-Header finden sich neben den Standard-Einträgen gemäß dem Internet-Standard RFC 5322 auch Nutzer-definierte Einträge, erkennbar am Präfix 'X-', die z.B. vom Email-Client-Programm eingefügt sein können. Der untere Teil der Email besteht aus einer weitergeleiteten Email, die durch eine Strich-Markierung eingeleitet wird. Der eigentliche Text, der dem Absender als Autor zugeordnet werden kann, besteht also aus nur einer Zeile (*Did you already ...*), dazu kommen die Grußformel (*Drew,*) und die Abschiedsformel (*Denise*).

Vor der eigentlichen Projektaufgabe stand eine Reihe von Vorverarbeitungen an: Der komplette Header wurde gelöscht, die Information über Absender und Adressat in anderer Form abgespeichert. Abschnitte, die aus anderen Mails in Form von Zitaten übernommen wurden, mussten erkannt und gelöscht werden (da sie von anderen Autoren stammen). Ebenso wurden Abschnitte, die aus weitergeleiteten Mails bestanden, entfernt (s. die Beispiel-Email in Abb. 2). Gegebenenfalls musste Markup, wie z.B. HTML-Tags, gelöscht werden. Alle verbleibende Information wurde in einem XML-Format abgelegt. Zuletzt wurden die Mails umgeordnet: statt nach ihrem Empfänger wurden sie nach ihrem Absender sortiert (da es im Projekt ja um die automatische Erkennung des Autors ging).

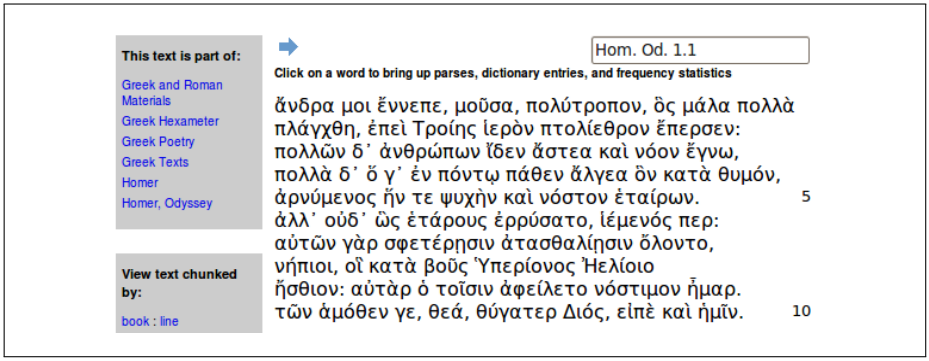
Als nächster Schritt stand die automatische Analyse mit Hilfe frei verfügbarer Tools an. Hierfür wählte die studentische Gruppe als POS-Tagger den Stanford-Tagger und als Parser RASP.[15] Der ursprüngliche Plan sah vor, ausgewählte Merkmale, die aus den Analysen des Taggers und Parsers extrahiert werden sollten, mit WEKA[16] weiter zu verarbeiten, einem Tool, das eine Reihe von Algorithmen für *data mining* für die Anwendung bereit stellt. Allerdings erwies sich die Datenmenge als zu groß für eine Verarbeitung durch den Tagger und Parser (im Rahmen des Kurses). Letztlich wurde der Klassifikator daher auf oberflächennahen Merkmalen trainiert: Anzahl der Sätze in der Email, durchschnittliche Satzlänge, Type-Token-Ratio, Grußformel (abstrahiert), prozentualer Anteil an Satzzeichen. Als besonders distinktives Merkmal stellte sich die Grußformel heraus.

Beispiel-Korpus: Homer Eine zweite Gruppe von Studenten wählte als Textgrundlage die Ilias und Odyssee von Homer. In der *Homerischen Frage* geht es darum, ob die beiden Epen von nur einem Autor geschrieben wurden — eine bis heute nicht gelöste Frage. Laut dem *Unitarismus* gab es einen einzigen Dichter beider Werke. Im Gegensatz dazu besagt die *Oral-Poetry-Theorie*, dass die Epen zuerst mündlich improvisiert und tradiert wurden, bevor sie schriftlich festgehalten wurden. Nach der *Analyse*-Theorie wiederum gab es für beide Epen jeweils einen "Hauptdichter", dessen Werk im Laufe der Zeit durch mehrere "Nebendichter" ergänzt und angereichert wurde.

Die Studenten hatten entsprechend zwei Aufgaben: zum einen zu berechnen, wie homogen die beiden Epen in sich sind (also Teilabschnitte mit dem Gesamtepos zu vergleichen); zum anderen zu berechnen, wie ähnlich die beiden Epen zueinander sind (und das Ergebnis in Relation zu setzen zum ersten Ergebnis).

Die Texte Homers werden über das Projekt Perseus angeboten und sind (automatisch) annotiert mit POS und Morphologie.[17] Allerdings sind die Texte und Annotationen nicht direkt downloadbar, sondern nur über ein Web-Interface zugänglich, s. Abb. 3. Durch einen Mausklick auf ein Wort erhält man die zugehörige POS- und morphologische Analyse.

[15] Stanford Tagger: `http://nlp.stanford.edu/software/tagger.shtml`
RASP: `http://www.informatics.sussex.ac.uk/research/groups/nlp/rasp/`
[16] `http://www.cs.waikato.ac.nz/ml/weka/`
[17] `http://www.perseus.tufts.edu/hopper/`




```
<a href="morph?l=a%29%2Fndra&amp;la=greek">ἄνδρα</a>
<a href="morph?l=moi&amp;la=greek&amp;prior=a)/ndra">μοι</a>
<a href="morph?l=e%29%2Fnnepe&amp;la=greek&amp;prior=moi">ἔννεπε</a>,
<a href="morph?l=mou%3Dsa&amp;la=greek&amp;prior=e)/nnepe">μοῦσα</a>,
<a href="morph?l=polu%2Ftropon&amp;la=greek&amp;prior=mou=sa">πολύτροπον</a>,
<a href="morph?l=o%28%5Cs&amp;la=greek&amp;prior=polu/tropon">ὃς</a>
<a href="morph?l=ma%2Fla&amp;la=greek&amp;prior=o(\s">μάλα</a>
<a href="morph?l=polla%5C&amp;la=greek&amp;prior=ma/la">πολλὰ</a><br />
```

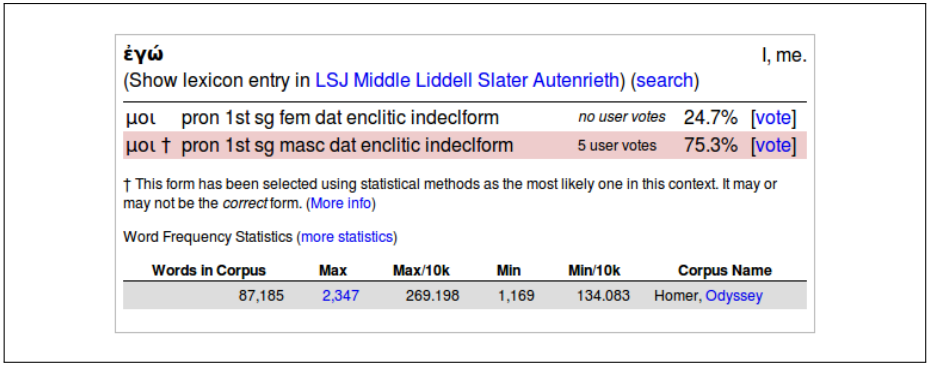


Abbildung 3: Screenshots vom Perseus-Projekt: (i) Oben ist der Beginn der Odyssee von Homer abgedruckt. (ii) In der mittleren Box steht die (verkürzte) HTML-Darstellung der ersten Verszeile. Aus dem 'href'-Attribut wird ersichtlich, dass für die morphologische Analyse, die durch Mausklick abgerufen wird, intern Betacode benutzt wird (z.B. ist '%29%2F' die Hexadezimal-Darstellung von ')/' und steht für die altgriechischen Diakritika über dem ersten Buchstaben im Text). Außerdem kann man sehen, dass für die morphologische Analyse das jeweilige Vorgängerwort ('prior') mitgenutzt wird. (iii) Unten wird der Lexikoneintrag zum zweiten Wort gezeigt, dem Personalpronomen *μοι* 'mir', mit POS- und morphologischer Analyse. Die statistische Disambiguierung präferiert hier die maskuline Lesart des Pronomens.

Mit Hilfe von Download-Programmen wie *wget* konnte die Projektgruppe den vollständigen Text der Ilias und der Odyssee (in Betacode) herunterladen. Eine Nutzung des Perseus-Tools für die POS- und Morphologie-Analyse war jedoch losgelöst vom Web-Interface nicht möglich, so dass letztlich das heruntergeladene Korpus nicht automatisch annotiert werden konnte. Statt linguistische Merkmale zu nutzen, musste die Projektgruppe daher doch auf einen *bag of words*-Ansatz zurückgreifen.

3.2 Korpora für den Ressourcenaufbau

Im zweiten Beispiel aus der Computerlinguistik werden Korpora für den Aufbau von Sprachressourcen genutzt. Konkrete Aufgabe für die Studenten war, mit Hilfe eines Korpus ein deutsches Nomenlexikon mit Angabe der Flexionsklasse zu erstellen.

Die Aufgabe wurde von zwei Gruppen bearbeitet, die mit unterschiedlichen Ressourcen arbeiteten: Gruppe 1 verwendete ein deutsches Zeitungskorpus, das mit Hilfe des RFTaggers[18] automatisch annotiert worden war, und zwar mit POS-, Lemma- und morphologischer Information. Hier bestand die Aufgabe darin, aus Tupeln bestehend aus einem Lemma, den zugehörigen Wortformen und der Flexionsinformation die damit kompatible Flexionsklasse zu bestimmen. Für die Fälle, in denen nicht genügend Belege für die verschiedenen Flexionsformen gefunden wurden, mussten Heuristiken entwickelt werden, um die wahrscheinlichste Klasse zu bestimmen. Außerdem waren Lösungen zu entwickeln für homographe Lemmata mit unterschiedlichen Flexionsklassen.[19]

Gruppe 2 verwendete dasselbe Korpus, aber nutzte neben den Wortformen nur die POS-Annotation. Hier musste also zusätzlich eine Methode entwickelt werden, zusammengehörige Wortformen als solche zu erkennen. Mit diesem Ansatz konnten dann auch (Vorschläge für) Lexikoneinträge von Nomen, die vom RFTagger nicht lemmatisiert wurden, erzeugt werden — das waren im betreffenden Korpus immerhin rund 20% aller Nomen (11% der normalen Nomen, 45% der Eigennamen).

Als Fazit aus den Szenarien in diesem Abschnitt lässt sich festhalten, dass es für die computerlinguistische Lehre essenziell ist, dass die Korpora frei zum Download und zur lokalen Weiterverarbeitung zur Verfügung stehen. Außerdem sollten Analysetools ebenfalls für den lokalen Einsatz genutzt werden können. Alternativ könnten (frei verfügbare) annotierte Korpora dazu verwendet werden, eigene Tools zu trainieren.

Während im ersten Szenario die Textauswahl eine große Rolle für die Aufgabe spielte, war die Textgrundlage im zweiten Szenario sekundär. Hier kam es vorwiegend darauf an, dass die Texte mit gängigen Analysetools mit zufriedenstellender Performanz automatisch annotiert werden konnten.

[18] `http://www.ims.uni-stuttgart.de/projekte/corplex/RFTagger/`

[19] Z.B. lautet für das Lemma *Bank* die Form des Nominativ Plural in der Bedeutung als Geldinstitut *Banken*, in der Bedeutung als Sitzmöbel *Bänke*.

4 Schluss

In den vorgestellten Szenarien aus der Lehre spielen Korpora eine unterschiedliche Rolle: In der Theoretischen Linguistik dienen sie einerseits als Belegsammlung, aus der interessante Einzelbelege mit Hilfe der Annotationen und geeigneter Suchanfragen gezielt herausgesucht werden können. Andererseits werden Korpora als Gegenstand für die eigene (manuelle) Analyse und Annotation gesehen; die relevanten Merkmale sind hier noch nicht (vollständig) annotiert. In beiden Fällen geht es darum, Korpora zur Entwicklung und Validierung linguistischer Theorien zu nutzen.

Die Computerlinguistik kann in ähnlicher Weise auf bereits annotierte Korpora zurückgreifen und ausgewählte Merkmale für die (automatische) Analyse nutzen. Häufig sind geeignete Korpora allerdings nicht oder nicht frei verfügbar. Dann ist es Teil der Aufgabe, Korpora selbst aufzubereiten und automatisch zu annotieren.

Aus den Szenarien ergeben sich unterschiedliche Anforderungen an die Sprachressourcen. Auf der einen Seite sollen technische Details und Hürden (wie z.B. die interne XML-Kodierung der Daten oder die Tool-Installation) in der Lehre für theoretische Linguisten ausgeklammert werden können. Daher plädiere ich dafür, dass Korpora und Annotationstools (für die manuelle Annotation) über das Web benutzbar sein sollen. Wichtig wäre ein Anschluss der Annotationstools an Korpus-Suchtools.

Auf der anderen Seite benötigt man in der Lehre für die Computerlinguistik den vollen Zugriff auf die Ressourcen. Für viele Anwendungen werden zudem größere Datenmengen benötigt, was allerdings dank freier Quellen wie Wikipedia für viele Sprachen kein Problem mehr ist. Freie Tools für automatisches POS-Tagging gibt es (mit Sprachmodellen) für "gängige" Sprachen wie Englisch oder Deutsch in genügender Anzahl. Tools für das Parsing oder Chunking hingegen sind seltener und oft nicht effizient genug.

Bei der Arbeit mit "echten" Daten, in realistischen Szenarien, werden die Computerlinguistik-Studenten früh mit Problemen wie dem Daten-Encoding oder der Datengröße konfrontiert, die mit Sicherheit auch in ihrem späteren Berufsleben eine Rolle spielen werden.

Zuletzt ein Vergleich zum Verhältnis zwischen dem Gegenstand *Text* und dem daraus gewonnenen Wissen in den verschiedenen Szenarien. Texte stellen linguistisches Wissen in *extensionaler* Weise dar, dagegen haben linguistische Hypothesen und Theorien den Anspruch, linguistisches Wissen in *intensionaler* Weise abzubilden. Anders formuliert könnte man sagen, dass Texte die "Produkte angewandten linguistischen Wissens" sind, während linguistische Theorien die zugrunde liegenden Regeln und Generalisierungen, d.h. das linguistische Wissen selbst erfassen wollen.

Aufgabe der Theoretischen Linguistik ist es nun, aus dem Text, d.h. aus den "beobachtbaren" sprachlichen Daten, auf die dahinter liegenden Regularitäten und Gesetzmäßigkeiten rückzuschließen. Auf ganz ähnliche Art versuchen Computerlinguisten, mit statistischen Methoden geeignete Sprachmodelle zu erstellen, die die beobachteten Daten möglichst optimal erklären.

In beiden "Lagern", der Theoretischen Linguistik wie der Computerlinguistik, wird der Weg vom extensionalen Objekt (dem Text) zum intensionalen Objekt (der Theorie/dem Modell) dadurch vereinfacht, dass geeignete *Abstraktionen* vorgenommen werden: Durch manuelle oder automatische Annotationen (POS-Tagging und andere Analyseschritte) werden sprachliche Daten *reduziert* auf die für die Theorie- und Modellbildung relevanten Eigenschaften und vereinfachen dadurch das Aufspüren von Regularitäten. In diesem Abstraktionsschritt werden die Daten allerdings nicht nur reduziert, sondern gleichzeitig auch *angereichert*: Durch die Abstraktion wird nämlich sprachliches Wissen, das im Text nur implizit (nämlich extensional) ausgedrückt ist, *explizit* gemacht.

Literatur

Brants, T. (1997). The NeGra export format. CLAUS Report Nr. 98. Universität des Saarlandes, Computerlinguistik, Saarbrücken.

Cook, P. and Bildhauer, F. (2011). Annotating information structure: The case of topic. In Dipper, S. and Zinsmeister, H., editors, *Proceedings of the Workshop 'Beyond Semantics: Corpus-based Investigations of Pragmatic and Discourse Phenomena*, volume 3 of *Bochumer Linguistische Arbeitsberichte (BLA)*, pages 45–56.

Filippova, K. and Strube, M. (2007). German Vorfeld and local coherence. *Journal of Logic, Language, and Information (JoLLI)*, 16(4):465–485. Special Issue on Coherence in Dialogue and Generation.

Frey, W. (2004). A medial topic position for German. *Linguistische Berichte*, 198:153–190.

Götze, M., Weskott, T., Endriss, C., Fiedler, I., Hinterwimmer, S., Petrova, S., Schwarz, A., Skopeteas, S., and Stoel, R. (2007). Information structure. In Dipper, S., Götze, M., and Skopeteas, S., editors, *Information Structure in Cross-linguistic Corpora: Annotation Guidelines for Phonology, Morphology, Syntax, Semantics, and Information Structure*, volume 7 of *Working Papers of the SFB632, Interdisciplinary Studies on Information Structure (ISIS)*, pages 147–187. Universitätsverlag Potsdam.

Klimt, B. and Yang, Y. (2004). Introducing the Enron corpus. In *Proceedings of the Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS-2004)*.

Mengel, A. and Lezius, W. (2000). An XML-based encoding format for syntactically annotated corpora. In *Proceedings of the Second International Conference on Language Resources and Engineering (LREC)*, pages 121–126.

Schiller, A., Teufel, S., Stöckert, C., and Thielen, C. (1999). Guidelines für das Tagging deutscher Textcorpora mit STTS (kleines und großes Tagset). Technical report, Universitäten Stuttgart und Tübingen, `http://www.ims.uni-stuttgart.de/projekte/corplex/TagSets/stts-1999.pdf`.

Shetty, J. and Adibi, J. (2004). The Enron email dataset: Database schema and brief statistical report. Technical report, Information Sciences Institute. `http://www.isi.edu/~adibi/Enron/Enron_Dataset_Report.pdf`.

Speyer, A. (2007). Die Bedeutung der Centering Theory für Fragen der Vorfeldbesetzung im Deutschen. *Zeitschrift für Sprachwissenschaft*, 26:83–115.

Alexander Mehler, Silke Schwandt, Rüdiger Gleim, Bernhard Jussen

Der eHumanities Desktop als Werkzeug in der historischen Semantik: Funktionsspektrum und Einsatzszenarien

Die *Digital Humanities* bzw. die *Computational Humanities* entwickeln sich zu eigenständigen Disziplinen an der Nahtstelle von Geisteswissenschaft und Informatik. Diese Entwicklung betrifft zunehmend auch die Lehre im Bereich der geisteswissenschaftlichen Fachinformatik. In diesem Beitrag thematisieren wir den *eHumanities Desktop* als ein Werkzeug für diesen Bereich der Lehre. Dabei geht es genauer um einen Brückenschlag zwischen Geschichtswissenschaft und Informatik: Am Beispiel der historischen Semantik stellen wir drei Lehrszenarien vor, in denen der *eHumanities Desktop* in der geschichtswissenschaftlichen Lehre zum Einsatz kommt. Der Beitrag schließt mit einer Anforderungsanalyse an zukünftige Entwicklungen in diesem Bereich.

1 Einleitung

In einer Zeit aufkommender Bestrebungen um die Entwicklung einer *Digital Humanities* als wissenschaftliche Disziplin mit eigenem Erkenntnisinteresse und eigenen Untersuchungsmethoden nimmt es nicht Wunder, dass die Ergebnisse dieser Arbeit zunehmend auch im Bereich der Lehre erprobt werden. In diesen Kontext ordnen wir ähnlichlautende Bestrebungen um die Entwicklung des *Computing in the Humanities* bzw. der *Computational Humanities* ein. Der Unterschied dieser Bestrebungen resultiert aus einer stärkeren Betonung der computerbasierten Repräsentation geisteswissenschaftlicher Interpretationsstrukturen als zentrale Aufgabe der *Digital Humanities* gegenüber einer stärkeren Betonung der Algorithmisierung von Explorationen solcher Strukturen als zentrale Aufgabe der *Computational Humanities* (vgl. hierzu den Workshop von Heyer and Büchler, 2010). Während also geisteswissenschaftlich ausgerichtete Informatiker im Bereich der *Digital Humanities* Interpretationsresultate computerbasiert modellieren und verarbeiten, sind es Informatiker im Bereich der *Computational Humanities*, die auf die automatische Exploration solcher Interpretationsresultate zielen. Wegen des Fokus auf lehrorientierte Nutzungsszenarien des vorliegenden Themenhefts vernachlässigen wir diesen Unterschied im Rahmen dieses Beitrags und sprechen unterschiedslos von geisteswissenschaftlicher Fachinformatik, um diese beiden Bereiche zu denotieren.

Gemäß dem interdisziplinären Grundcharakter all dieser Ansätze steht ebenso außer Frage, dass die geisteswissenschaftliche Fachinformatik nicht allein die Lehre in der Informatik selbst betrifft. Vielmehr geht es darum, lehrorientierte Anwendungsfelder

insbesondere in den geisteswissenschaftlichen Einzeldisziplinen aufzutun. Hierzu zählen die sprachwissenschaftlichen Disziplinen, die infolge ihres frühen Bezugs zur Computerlinguistik längst im Austausch mit der Informatik stehen, ebenso wie jene geisteswissenschaftlichen Disziplinen, für welche die Anwendung computerbasierter Methoden nach wie vor eine Seltenheit darstellt. Zum gegenwärtigen Zeitpunkt ist die Geschichtswissenschaft weitgehend letzterem dieser beiden Lager zuzurechnen.

In diesem Kontext thematisiert das vorliegende Papier einen Brückenschlag ausgehend von der Informatik hin zur textorientierten Geschichtswissenschaft und umgekehrt. Damit bleibt die Anwendbarkeit des zu umreißenden Instrumentariums auch in anderen Disziplinen unbestritten.[1] Der Erfahrungshintergrund der Lehre in der geisteswissenschaftlichen Fachinformatik, der hier skizziert werden soll, betrifft jedoch vorrangig die Geschichtswissenschaft.

Allgemein gesprochen stehen wir bei diesem Brückenschlag vor einer Aufgabe, die in Abbildung 1 ihren schematisierten Niederschlag findet und wie folgt umschrieben werden kann:

1. Das eine Extrem bildet jene Auffassung von Informatik, für welche die Orientierung an formalisierbaren Repräsentationen und berechenbaren Algorithmen zur Erfassung von Strukturen und Prozessen des jeweiligen Gegenstandsbereichs kennzeichnend ist. Was nicht algorithmisierbar oder gar nicht formalisierbar ist, steht außerhalb dieser Auffassung des Gegenstandsbereichs der Informatik – dieser Lesart entspricht in etwa die klassische Auffassung der Kognitionswissenschaft bei Fodor and Pylyshyn (1988). Im Idealfall, sozusagen fernab von der Komplexität semiotischer Prozesse (Petitot et al., 2000; Rieger, 2003), ist die Informatik allein mit vollständig automatisierbaren Prozessen befasst, deren Qualitätsmessung strikten Regeln folgt und also nach dem Muster überwachter Lernprozesse abläuft.

2. Das andere Extrem bildet jene Auffassung von Geisteswissenschaft diesseits der Anwendung computerbasierter Methoden, die sich am Ideal einer hermeneutischen, rein intellektuellen, verstehensgeleiteten Wissenschaft orientiert, deren Erklärungsmodelle wegen ihres konstruktivistischen Gehalts im Extremfall nicht erschöpfend reproduzierbar sind. Die Aktualisierung solcher Modelle setzt vielmehr jene Art von Interpreten als unabdingbares Aktualisierungsmedium voraus, das keines seiner Verstehensschritte je an eine Maschine delegiert noch delegieren könnte.

3. In diesem Spannungsfeld ist die geisteswissenschaftliche Fachinformatik nun – vergleichbar mit der Computersemiotik (Rieger, 1989; Andersen, 1990) – als der Versuch aufzufassen, eine Position zwischen diesen Extremen einzunehmen, und zwar entlang zweier Orientierungen: dies betrifft zum einen die Annahme von der Formalisierbarkeit und Algorithmisierbarkeit nicht trivialer, erkenntnisfördernder Teile semiotischer Prozesse sowie die Forderung nach der Interpretierbarkeit

[1] Siehe beispielsweise Mehler et al. (2011) für eine Anwendung im Bereich der Romanistik oder Gleim et al. (2010) für eine Anwendung im Bereich der Bildwissenschaft.

> jeglicher Formalisierungs- und Algorithmisierungsresultate, und zwar so, dass diese an die hermeneutische Tradition anknüpfbar sind. Einer solcherart verstandenen geisteswissenschaftlichen Fachinformatik geht es folglich darum, den Bereich jener Methoden zu vergrößern, die beides zugleich sind: formalisier- und berechenbar wie auch geisteswissenschaftlich interpretierbar.

Offenbar adressiert die Umsetzung letzteren Programms eine Art von Algorithmus, der – im Sinne des *Human Computation* (von Ahn, 2008) – ein enges Wechselspiel von Automatisierung und geisteswissenschaftlicher Interpretation voraussetzt – fernab vom Ideal überwachten Lernens, aber auch fernab vom Paradigma des unüberwachten Lernens insofern dieses lediglich auf dem Mechanismus des *Relevance Feedback* beruht.[2]

Diesem Ansatz gemäß ist die geisteswissenschaftliche Fachinformatik genauso wenig eine Hilfswissenschaft irgendeiner Geisteswissenschaft wie auch umgekehrt geisteswissenschaftliche Objekte nicht einfach den Gegenstandsbereich der ansonsten unverändert bleibenden Informatik erweitern. Vielmehr ist die Vorstellung leitend, dass die Anwendung informationswissenschaftlicher Methoden nur dann *Humanities* als *Computational* kennzeichnet, wenn sie den Erkenntnishorizont der betroffenen Geisteswissenschaften erweitert, ohne bereits mit ihrem herkömmlichen, nicht-informationswissenschaftlichen Apparat erbringbar zu sein. Es geht also um eine Art von geisteswissenschaftlicher Disziplin, deren Gegenstandsbereich und Erkenntnisinteresse durch computerbasierte Methoden überhaupt erst ihre Prägung erhalten, und zwar jenseits etablierter Geisteswissenschaften. Als ein Beispiel für eine solche Blickrichtung können intertextuelle Strukturen gelten, die allein an solchen Textmengen beobachtbar sind, deren Größe anders als computerbasiert nicht zu bewältigen ist. In diesem Beispiel wird Intertextualität als Untersuchungsobjekt erst fassbar durch die computerbasierte Modellierung – und zwar in qualitativer wie auch quantitativer Hinsicht (Wagner et al., 2009).

Blickrichtungen dieser Art ist der eHumanities Desktop und seine Anwendung in Forschung und Lehre der Geschichtswissenschaft geschuldet. Im Folgenden wird dies an der historischen Semantik und ihrer Vermittlung in der Lehre exemplifiziert. Das Papier ist wie folgt organisiert: Sektion 2 beschreibt die historische Semantik als ein Teilgebiet der Geschichtswissenschaft. Dies betrifft ihr Erkenntnisinteresse ebenso wie ihre vorherrschende korpusbasierte Methodik. Sektion 3 erläutert eine Auswahl des Funktionsspektrums des eHumanities Desktops, jenes Systems also, das die Geschichtswissenschaft an der Goethe-Universität Frankfurt im Rahmen von drei Lehrszenarien eingesetzt hat bzw. einsetzt. Diese Szenarien werden in Sektion 4 beschrieben. Sektion 5 folgt mit einer Diskussion von Anforderungen, die sich aus diesen Erfahrungen an die Weiterentwicklung des eHumanities Desktop ergeben, ehe Sektion 6 mit einer Zusammenfassung und Schlussfolgerung schließt.

[2] Siehe Wagner et al. (2009) für einen solchen Ansatz im Bereich von *literary memory information systems*.

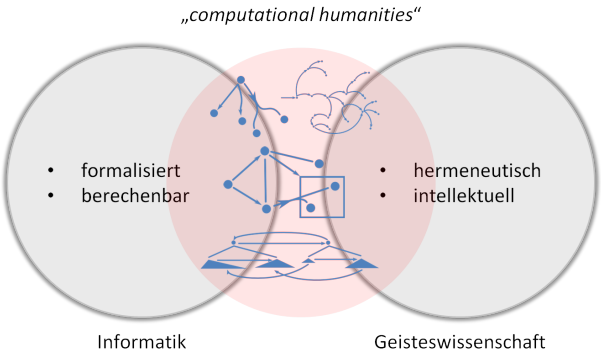


Abbildung 1: Die geisteswissenschaftliche Fachinformatik im Spannungsfeld von Informatik und Hermeneutik.

2 Historische Semantik: methodisch-theoretischer Hintergrund

In dieser Sektion skizzieren wir den erkenntnistheoretischen Rahmen, auf dem die Historische Semantik und damit letztlich die von Historikern angestrebte Nutzung des eHumanities Desktops ruht. Unter Verwendung der in den historischen Wissenschaften gängigen Fachterminologie wird dabei aufgezeigt, welchen Erkenntnisfortschritt der Einsatz des eHumanities Desktops – ob nun in Forschung oder Lehre – maßgeblich unterstützen soll. Dieser Fortschritt setzt an einer methodischen Verbindung von datengetriebener Korpuslinguistik, statistischer Textlinguistik und interpretatorischer Hermeneutik an, und macht somit die Zusammenarbeit von Historikern und Texttechnologen unabdingbar.

Die Historische Semantik widmet sich innerhalb der Geschichtswissenschaft der Erforschung von Bedeutung und insbesondere der Erforschung ihrer historischen Variabilität. Dabei nimmt sie verschiedene Medien und Objekte in den Blick, die sich zur Interpretation von Sinnzuschreibungsprozessen eignen. Hier geht es zunächst um eine textuell geprägte Historische Semantik.

Eine der Grundannahmen ist, dass Bedeutung jeweils im Gebrauch entsteht und immer wieder aktualisiert werden muss. So ist Bedeutung vor allem auch situativ variabel. Der Blick der Forschung richtet sich daher nicht allein auf bestehende Verknüpfungen von Zeichen und Sinn, sondern auf die Mechanismen der Sinnzuschreibung – Rieger (2003) spricht an dieser Stelle genauer von Prozessen der Bedeutungskonstitution. Diese gelten als von der Gesellschaft kontrollierbar, da sie es ist, die mit Hilfe von Sprache Sinn erzeugen muss. Es schließt sich die Annahme an, dass sprachliche Phänomene und Sinnzuschreibungsprozesse auch Aufschluss geben über ihre gesellschaftlichen Bedingungen. Eine der zentralen Fragen der Historischen Semantik ist die nach dem Verhältnis von Sprachwandel und gesellschaftlichen Wandlungsprozessen.

Die Geschichtswissenschaft kennt seit dem *linguistic turn* der 1970er Jahren verschiedene Traditionen, die sich mit der Erforschung von Sinn beschäftigt haben. Für

Deutschland ist in erster Linie die *Begriffsgeschichte* nach Reinhart Koselleck prägend gewesen. Koselleck versammelte eine Reihe von Kollegen um sich, mit denen er die *Geschichtlichen Grundbegriffe* schrieb, heute ein Standardwerk in der Geschichtswissenschaft. In Form von Handbucheinträgen widmeten sich die Autoren der politisch-sozialen Sprache, wie Koselleck (1972, XIII) sagt, den "Leitbegriffe[n] der geschichtlichen Bewegung, die, in der Folge der Zeiten, den Gegenstand der historischen Forschung ausmacht." Im Zentrum stehen dabei solche Begriffe, die aus Sicht des modernen Wissenschaftlers/der modernen Wissenschaftlerin gesellschaftliche Relevanz beanspruchen, und deren Geschichte. Dabei wird die Untersuchung des Begriffs immer rückgebunden an die sozialen Gegebenheiten einer Zeit. Daraus ergibt sich ein Geflecht aus Sprache und ihren gesellschaftlichen Bedingungen sowie, anders herum, aus gesellschaftlichen Strukturen und ihren sprachlichen Konstruktionen und Modalitäten.

Um den Blick der Moderne auf die Geschichte ihrer Begriffe zu lenken, hat die Historische Semantik ihre Methoden immer weiter verfeinert und auch quantitative Vorgehensweisen in ihr Konzept integriert. Ziel war die Abkehr von der Untersuchung verfassungsgeschichtlicher Termini und deren Geschichte, und zwar deutlicher als dies in den Anfängen der *Begriffsgeschichte* geschehen ist. Außerdem versprach der quantitative Zugriff eine Verbesserung der Materialbasis für sprachliche Analysen wie er von den Kritikern der Kosellecksschen Schule und auch in der *Conceptual History* (vgl. Skinner, 2003) gefordert wird. Ausgehend von den Forschungsinteressen der Pioniere der Historischen Semantik – oder der Methoden zur sprachlichen Analyse gegenwärtiger und vergangener Wirklichkeitskonstruktionen – haben sich entsprechende Untersuchungen lange Zeit auf Korpora der Moderne bzw. der Frühen Neuzeit konzentriert. Steinmetz (2008) vermittelt einen umfassenden Überblick zu dieser Forschung und deren Entwicklung in den letzten 40 Jahren. Auch wenn die *Geschichtlichen Grundbegriffe* in ihren Artikeln immer einen Blick auf die Vormoderne (Antike und Mittelalter) geworfen haben, blieb es doch lange bei eben diesem kurzen Blick.[3] Es sind aber gerade historische Langzeitkorpora, die im Rahmen historisch-semantischer Versuchsaufbauten besonders interessante Ergebnisse versprechen. Für Sprachwandeluntersuchungen bietet es sich beispielsweise an, diachron angelegte Textkorpora heranzuziehen, die für die Vormoderne wesentlich längere Zeiträume abdecken als für die Moderne. Vor allem die quantitative Analyse kann von der Untersuchung solcher Langzeitkorpora profitieren (Mehler et al., 2010, 2011). Sie gewinnt in diesem Rahmen eine breite Materialbasis.

Ausgehend von der *Begriffsgeschichte* können so semasiologische, diachron angelegte Untersuchungen der Verwendungsmuster bestimmter Wörter konzipiert werden. Der engere methodische Zugriff ist dabei eine quantifizierende Kookkurrenz-Untersuchung im Umfeld der Gebrauchssituationen der ausgewählten Vokabel. Innerhalb spezifischer Verwendungsmuster markieren die neben der Zielvokabel gebrauchten Wörter den situativ aktualisierten Sinn. Im Zeitalter der *Digital Humanities*, also der zunehmenden Verfügbarkeit großer Textvolumina in digitaler Form, steigen auch die Möglichkeiten computergestützter, korpusorientierter Methoden. Für die serielle Untersuchung

[3] Einen Überblick über semantische Studien für die Vormoderne bietet Rexroth (2009).

von Gebrauchskontexten bestimmter Vokabeln oder anders gestalteter sprachlicher Muster können so Quantifizierungen im großen Stil erstellt werden.[4] Korpora, die für solche Studien herangezogen werden können, liegen quer zu den vorhandenen digitalen Editionen mittelalterlicher Texte wie etwa der *Patrologia Latina* (PL) (Migne, 1855), der *digitalen Monumenta Germaniae Historica* (dMGH) bzw. der *elektronischen Monumenta Germaniae Historica* (eMGH) (Radl, 2007). Diese Sammlungen stellen die Ressourcen zur Verfügung, aus denen Korpora erstellt werden müssen, die sich an den spezifischen Anforderungen je verschiedener Forschungskonzepte und -fragen orientieren (vgl. Jussen 2006). Ziel ist es, aus den Gebrauchssituationen bestimmter Sprachmuster eine "kollektive Haltung" jenseits einzelner Autoren und ihrer Texte zu erarbeiten. Um diese dann in diachroner Perspektive betrachten zu können und sie gleichzeitig mit den Hypothesen der politischen Ideengeschichte abzugleichen, ist der Untersuchung ein Textkorpus zu Grunde zu legen, das es ermöglicht, homogene Beobachtungsstrecken über längere Zeiträume zu erstellen (Jussen, 2006).[5] Ein solches Korpus besteht vorzugsweise aus einzelnen Werken einer bestimmten Textsorte.[6] Rieger (1989) spricht in diesem Zusammenhang von dem Kriterium der pragmatischen Homogenität der den Gebrauchsanalysen zugrundezulegenden Korpora.

Neben der Entwicklung neuer Methoden und Werkzeuge bleibt es eine Herausforderung für die Historische Semantik, ihre Ergebnisse zu kontextualisieren. Für diese Aufgabe werden Theorien herangezogen, die sich dem Zusammenspiel von Sprache und Sozialstruktur widmen (vgl. Luhmann, 1993). Semantik kann so verstanden werden als "Vorrat an bereitgehaltenen Sinnverarbeitungsregeln" und die Gesellschaftsstruktur als Rahmen, der die Beliebigkeit dieses Vorrats für die Verarbeitung von Sinn einschränkt (Luhmann, 1993, 19). Semantik liefert also ein breites Spektrum an möglichen Regularitäten zur Verarbeitung für Umstände, Situationen und Handlungszusammenhänge, denen die Gesellschaft Sinn zuschreibt. Eine Untersuchung dieser sprachlich verfassten Phänomene verspricht neue Einsichten in die Strukturen historischer Gesellschaften. Für die Geschichtswissenschaft ergeben sich daraus Beschreibungsmuster und Deutungsschemata für die Vergangenheit, die in Dialog treten können mit gegenwärtigen Mechanismen der Weltdeutung.

3 Zum Funktionsspektrum des eHumanities Desktops

Der erkenntnistheoretisch-methodologische Grundriss, den die vorangehende Sektion von der historischen Semantik gegeben hat, steht aus computerlinguistischer Sicht in enger

[4] Vgl. zur Darstellung der technologischen Anforderungen Jussen et al. (2007) und Mehler et al. (2010).

[5] In einem zweiten Schritt würden diese dann untereinander verglichen, um textsortenspezifische Semantiken herausarbeiten zu können, die wiederum dem Konzept der "languages" in der *Conceptual History* ähneln.

[6] Im Rahmen des LOEWE-Schwerpunkt *Digital Humanities* (`http://digital-humanities-hessen.de`) wird in Kooperation mit dem BMBF-Projekts *Linguistic Networks* (`http://project.linguistic-networks.net`) derzeit der Versuch unternommen, alle Texte der Patrologia Latina nach Textsorten zu kategorisieren.

Nachbarschaft zur explorativen Korpuslinguistik (Rieger, 1998; Heyer et al., 2006; Evert, 2008). Folgerichtig ist das Methodenspektrum zur Bedienung dieses Erkenntnisinteresses durch kollokationsstatistische Methoden vorgezeichnet. In diesem Kontext beschreiben wir nun eine Auswahl von Funktionalitäten, welche der eHumanities Desktop für Forschung und Lehre bereitstellt (Gleim et al., 2009a,b, 2010): Ausgehend von dem zentralen Administrationsmodul des eHumanities Desktops (Sektion 3.1) betrifft dies den *Corpus Manager* (Sektion 3.2) für die Ressourcenverwaltung, das frei konfigurierbare Modul für die Annotation Ressourcen-orientierter Metadaten (Sektion 3.3), das zentrale Vorverarbeitungstool des Desktops (Sektion 3.4), das HSCM-Modul (Sektion 3.5), den *Lexicon Browser* (Sektion 3.6) als zentrale Schnittstelle für das Lexikon-orientierte Retrieval und schließlich das *Classifier Builder* genannte Modul, mit Hilfe dessen vektorbasierte SVM-Modelle gebaut werden können. Nachfolgend werden diese Module erläutert.

3.1 Das Administrator-Modul

Der produktive Einsatz eines Systems zur kollaborativen Korpusverwaltung und -analyse erfordert eine entsprechende Rechteverwaltung. Diese soll den Anforderungen einzelner Arbeitsgruppen bis hin zu verteilten Forschungsprojekten gerecht werden und die Kollaborationsstrukturen solcher Projekte widerspiegeln. Abbildung 2 zeigt schematisch die Stammdatenverwaltung des eHumanities Desktop. Grundsätzlich wird unterschieden zwischen Autoritäten und den Ressourcen, hinsichtlich derer diese Berechtigungen besitzen. Eine Autorität ist zunächst ein abstrakter Begriff, der Benutzer und Benutzergruppen umfasst. Ein Benutzer kann beliebig vielen Gruppen angehören und erhält dadurch deren Berechtigungen. Für jede Benutzergruppe ist ein Eigentümer definiert (z.B. der Projektleiter einer Gruppe), welcher die Mitgliedschaften unabhängig vom zentralen Administrator verwalten kann.

Eine Ressource ist wiederum ein Sammelbegriff, der vor allem Repositorien und Dokumente umfasst. Die Berechtigungen auf Ressourcen können abgestuft gesetzt werden, ausgehend von reinen Leserechten, über Schreib- und Löschberechtigungen bis hin zu Freigabeberechtigungen. In der Praxis bedeutet dies, dass zum Beispiel Mitglieder einer Arbeitsgruppe Textsammlungen aufbauen und bearbeiten können und im Rahmen von Kooperationen die Möglichkeit erhalten, Projektpartnern Leseberechtigung auf den Daten zu vergeben. Diese Zugriffskontrolle ist vor allem für die Arbeit mit urheberrechtlich geschütztem Material relevant. Darüber hinaus besteht die Möglichkeit, nicht nur Dokumente und Repositorien, sondern auch Programmfunktionen freizugeben bzw. zu sperren. Auf diese Weise kann einer Anwenderin bzw. einem Anwender gezielt der Funktionsumfang bereitgestellt werden, der für ihre bzw. seine Arbeit relevant ist. Mit dem Anwendungsmodul *Administrator* steht somit ein Werkzeug zur Verfügung, Einzelbenutzer- und Benutzergruppenberechtigungen zu verwalten.

Genau diese Art der Berechtigungsverwaltung dient nicht nur zur Unterstützung von Forschungsprojekten, sondern erweist sich auch im Zusammenhang von Lehrveranstaltungen von großem Nutzen. Dies betrifft die Binnendifferenzierung der Zugänglichkeit einzel-

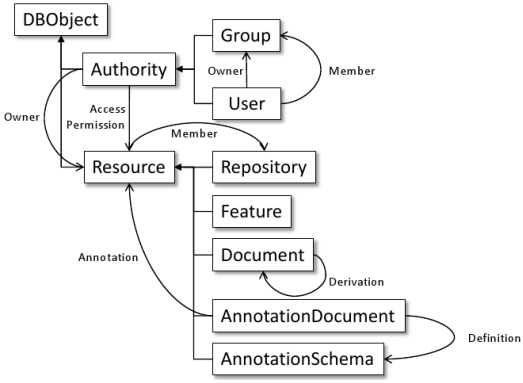


Abbildung 2: Das Modell der Stammdatenverwaltung des eHumanities Desktops.

ner Ressourcen ebenso wie die Freistellung oder Einschränkung von Einzelstudierenden- oder Studierendengruppen-bezogenen Funktionalitäten.

3.2 Das Corpus Manager-Modul

Der *Corpus Manager* ist ein zentrales Modul, das den webbasierten Aufbau von Korpora sowie die Verwaltung von Ressourcen ermöglicht. Dokumente können entweder direkt vom Arbeitsplatzrechner aus hochgeladen oder durch die Angabe einer URL aus dem Internet heruntergeladen werden. Umgekehrt ist es möglich, von jedem Rechner mit Internetanbindung über den Webbrowser auf sämtliche Daten und Anwendungen des eHumanities Desktop zuzugreifen. Der Desktop eignet sich damit insbesondere auch für den gruppenarbeitsorientierten Betrieb in ganzen Rechnerräumen.

Für viele Dokumenttypen und -arten stellt der Desktop Programme zur Betrachtung bereit, etwa für TEI P5, HTML und auch für Bilder. Der graphische Aufbau sowie die Benutzerführung orientiert sich dabei am Windows Explorer zur Dokumentverwaltung. Anders als etablierte Dateisysteme setzt der eHumanities Desktop jedoch nicht auf eine feste Verbindung von Dokumenten und Verzeichnissen. Vielmehr kann ein Dokument in verschiedenen Repositorien verwendet werden, ohne dass das Dokument dafür jedesmal kopiert werden muss. Das bietet den Vorteil, dass unterschiedliche Korpora auf der gleichen Dokumentbasis speichereffizient aufgebaut werden können. Wird ein Dokument bearbeitet oder mit Zusatzinformationen annotiert, so sind diese automatisch in allen Repositorien verfügbar, denen das betreffende Dokument angehört. Die Annotation von Dokumenten, welche im *Corpus Manager* durchgeführt werden kann, umfasst mehrere Arbeitsschritte, die im Folgenden skizziert werden.

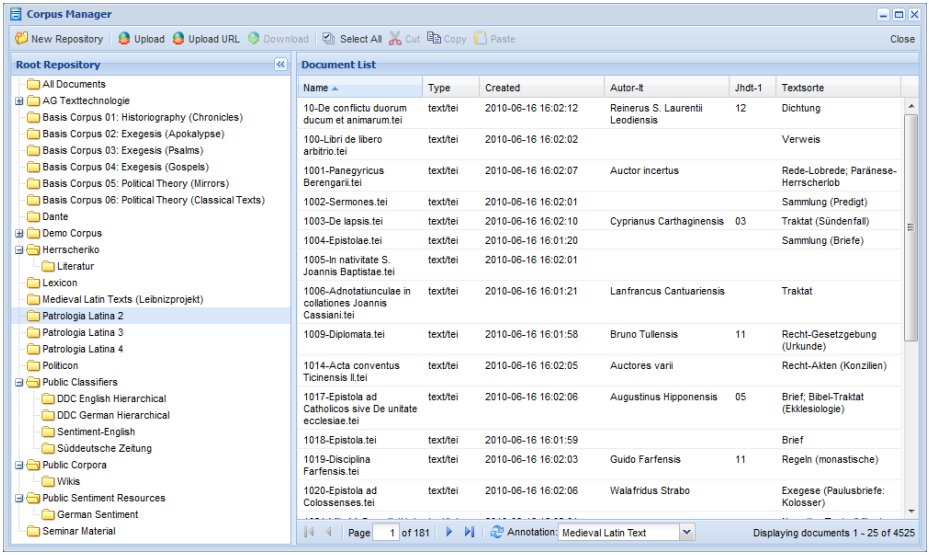


Abbildung 3: Ein Screenshot des *Corpus Manager* des *eHumanities Desktop*.

3.3 Das Annotator-Modul

Die Auszeichnung von Dokumenten mit Metadaten ist ein zentraler Baustein des Aufbaus und der anschließenden Analyse von Korpora. Der eHumanities Desktop trägt dieser Aufgabe durch Bereitstellung umfangreicher Funktionen Rechnung. So haben Anwender die Möglichkeit, selbstständig Datenfelder im Rahmen von Annotationsschemata zu definieren, um auf dieser Grundlage ihre Ressourcen auszuzeichnen. Es werden also keine Standardschemata vorgegeben. Nichtsdestotrotz reicht die Ausdrucksmächtigkeit des Systems dazu aus, gängige Standards, wie etwa *Dublin Core*[7], abzubilden.

Abbildung 4 zeigt einen Screenshot des *Schema Editor* des Desktops, in dem ein Datenfeld des Schemas "Medieval Latin Text" bearbeitet wird. Neben einem Namen kann unter anderem der Datentyp (Text, Zahl, Datum etc.) und dessen Wertebereich festgelegt werden. Abbildung 5 zeigt die Auszeichnung eines Dokuments auf der Grundlage dieses Schemas. Die Datenfelder von Annotationen können mittels des *Corpus Manager* auch in Tabellenform angezeigt, durchsucht und sortiert werden. Dies wird in Abbildung 3 exemplifiziert. Aus der Sicht der Datenmodellierung werden Annotationen von Ressourcen ebenfalls als Ressource betrachtet und fallen somit unter die Rechteverwaltung. Es steht damit jeder Anwenderin bzw. jedem Anwender frei, eine Ressource zu annotieren und diese Annotation mit anderen Anwendern zu teilen. Auf diese Weise

[7] `http://dublincore.org`

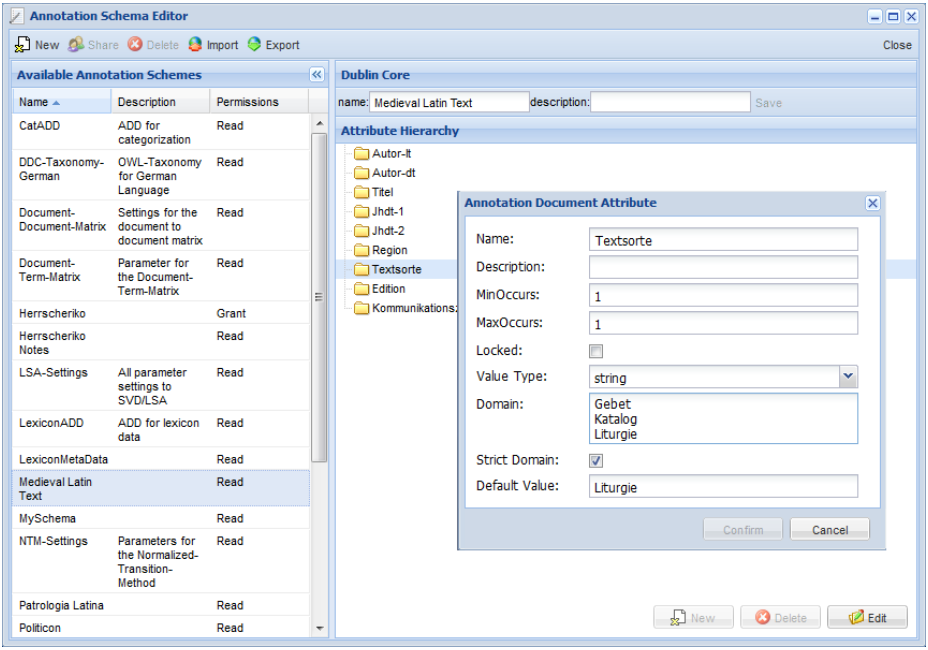


Abbildung 4: Ein Screenshot des *Annotation Schema Editor*.

wird eine kollaborative Dokumentauszeichnung unterstützt, was wiederum in der Lehre im Rahmen der kollaborativen Bearbeitung derselben Korpora eingesetzt werden kann.

3.4 Das Preprocessor-Modul

Für den sinnvollen Einsatz von Texten in korpuslinguistischen Studien ist eine linguistische Vorverarbeitung unabdingbar. Diese umfasst in der Regel die Tokenisierung, Lemmatisierung und Wortartenauszeichnung sowie die Erkennung logischer Dokumentstrukturen. Zur Repräsentation solcherart ausgezeichneter Dokumente verwendet der eHumanities Desktop den Dokumentrepräsentationsstandard der *Text Encoding Initiative*, das heißt TEI P5 (TEI Consortium, 2010). Darüber hinaus bietet das System mit dem *Preprocessor* ein Funktionsmodul, mit Hilfe dessen Anwender selbstständig Rohdaten (z.B. im PDF-, RTF- oder HTML-Format) vorverarbeiten und nach TEI P5 überführen können.

Aktuell wird die Verarbeitung von deutschen, englischen sowie von lateinischen Texten unterstützt (Waltinger, 2010; Mehler et al., 2008). Über die Benutzerschnittstelle können beliebig viele Eingangsdokumente im Stapelbetrieb automatisch verarbeitet werden.

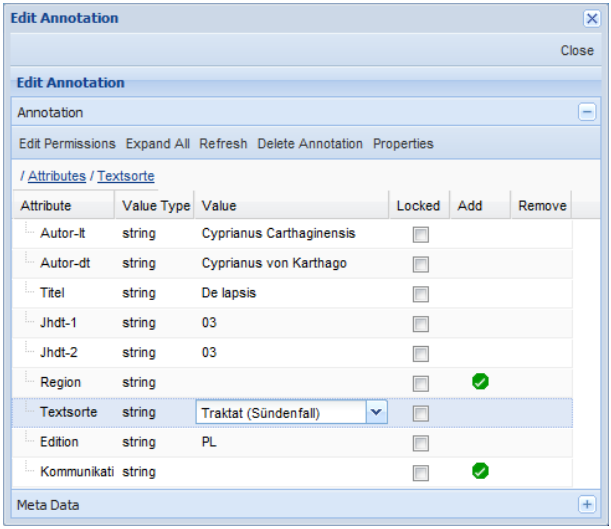


Abbildung 5: Ein Screenshot des ressourcenorientierten Annotationsdialogs.

Darüber hinaus können auch Texte direkt in ein Textfeld kopiert und vorverarbeitet werden. Abbildung 6 zeigt exemplarisch die Ausgabe einer solchen Verarbeitung in TEI P5.

3.5 Das Historical Semantics Corpus Management-System

Das *Historical Semantics Corpus Managements*-System (HSCM) zielt auf die texttechnologische Repräsentation und quantitative Analyse zeitlich geschichteter Corpora (Mehler et al., 2010). Es ermöglicht die Durchsuchung von Textsammlungen nach Belegstellen für einzelne Wörter oder Phrasen. Neben einer Auflistung von Texten, welche der Suchanfrage entsprechen, können gezielt Treffersätze und deren Kontext angezeigt werden. Die erste Version des HSCM war als eigenständige Client/Server Anwendung konzipiert und für die Verarbeitung der *Patrologia Latina* hin optimiert worden (Jussen et al., 2007). Der eHumanities Desktop inkludiert nunmehr die PL in einer Fassung, die aufgrund ihres Annotationsumfangs weit über die ursprüngliche SGML-basierte Ausgabe hinausgeht (Mehler et al., 2009). Die Dokumente dieses Korpus wurden mit einer Reihe von Metadaten ausgezeichnet, die insbesondere deren Entstehungszeitraum thematisieren. Mit der Integration in den eHumanities Desktop ist nun auch die Arbeit mit beliebigen, in TEI P5 repräsentierten Textsammlungen möglich. Die Syntax für die Suchanfragen ist intuitiv und leicht erlernbar. Intern werden die Anfragen in *Apache*

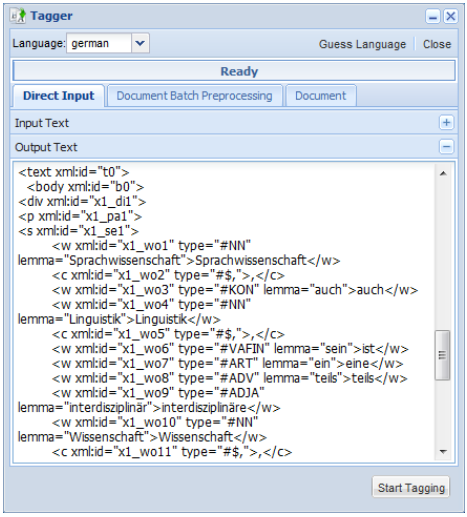


Abbildung 6: Ein Screenshot des *Preprocessor* in Zusammenhang mit einer Beispielausgabe basierend auf TEI P5.

Lucene[8] übersetzt. Die vormalige Verwendung einer XML-Datenbank hat sich zwischenzeitlich als deutlich weniger effizient erwiesen. Abbildung 7 zeigt einen Screenshot des HSCM mit den Ergebnissen einer Suchanfrage nach "Helmut Schmidt" in einem Korpus von Artikeln der "Süddeutschen Zeitung".

3.6 Der Lexicon Browser

Neben Texten und Bildern können auch Lexika als Dokumente im eHumanities Desktop einheitlich verwaltet werden. Der *Lexicon Browser* bietet eine Schnittstelle, die Anwender bei der Durchsuchung und Bearbeitung von Lexika unterstützt. Als Grundlage hierfür dient die einheitliche Repräsentation lexikalischer Ressourcen mit Hilfe des objektorientierten Datenmodells des eLexicon (Mehler et al., 2008). Für jedes Lexikon kann zudem ein Referenzkorpus angegeben werden, auf dessen Basis eine Reihe von Maßen zur Berechnung der Trennschärfe einzelner Lexikoneinträge berechnet wird. Dazu zählen die Frequenz, Textfrequenz und die inverse Dokumentfrequenz.

Da die Einträge der Lexika mit einer umfangreichen Ontologie von Wortarten ausgezeichnet sind, ermöglicht dies zusammen mit den Maßen eine differenzierte Suche nach Einträgen. Abbildung 8 zeigt einen Screenshot des Lexicon Browsers mit Ergebnissen einer Suchanfrage auf einem lateinischen Lexikon.

[8] Siehe `http://lucene.apache.org`

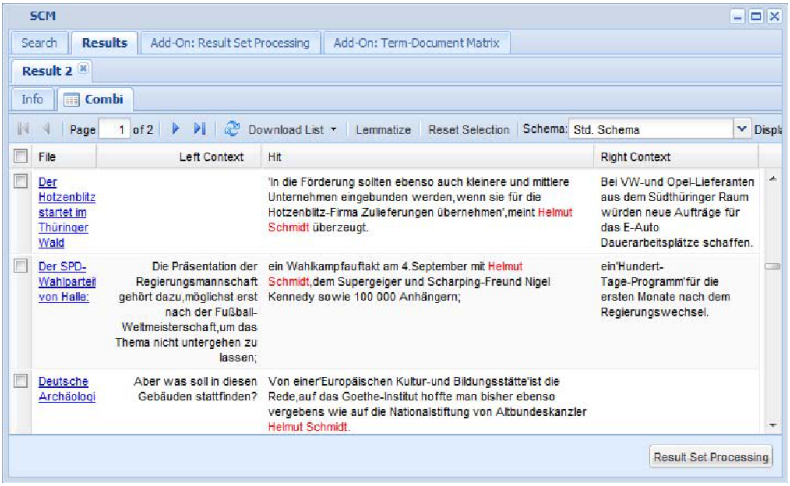


Abbildung 7: Ein Screenshot des HSCM mit den Ergebnissen einer Suchanfrage nach "Helmut Schmidt" im Korpus von Artikeln der Süddeutschen Zeitung.

3.7 Der Classifier Builder

Der *Categorizer* des eHumanities Desktop dient dazu, Textdokumente oder eingefügte Textausschnitte zu kategorisieren. Die Menge der hierfür verfügbaren Klassifikatoren ist nicht fest vorgegeben. Vielmehr können mit dem *Classifier Builder* eigene Klassifikatoren erstellt werden, um diese zur überwachten Textkategorisierung im Rahmen des Categorizer-Moduls heranzuziehen. Abbildung 9 exemplifiziert das Ergebnis der Kategorisierung eines Textausschnitts auf Basis der englischen *Dewey Decimal Classification* (DDC) (Waltinger et al., 2011). Die zutreffenden Kategorien sind nach ihrer Signifikanz absteigend sortiert und farblich abgesetzt.

4 Einsatzszenarien und Ziele in der Lehre

Nachdem nun der methodologische und erkenntnistheoretische Rahmen für Nutzung des eHumanities Desktops in der Geschichtswissenschaft skizziert wurde, geht es in dieser Sektion darum, Nutzungsszenarien des Desktops in der Lehre in der Geschichtswissenschaft zu beschreiben.

Die Historische Semantik ist an der Goethe-Universität bislang in drei Lehrformaten als Methoden- und Theorierahmen erprobt worden. Dabei standen ganz unterschiedliche Aspekte im Mittelpunkt. Dies betrifft eine Lehrveranstaltung zur "Politischen Sprache im Mittelalter" (durchgeführt von Bernhard Jussen und Silke Schwandt) wie auch, zweitens, eine Studiengruppe "Historische Semantik". Drittens sind mehrere "Anwender-

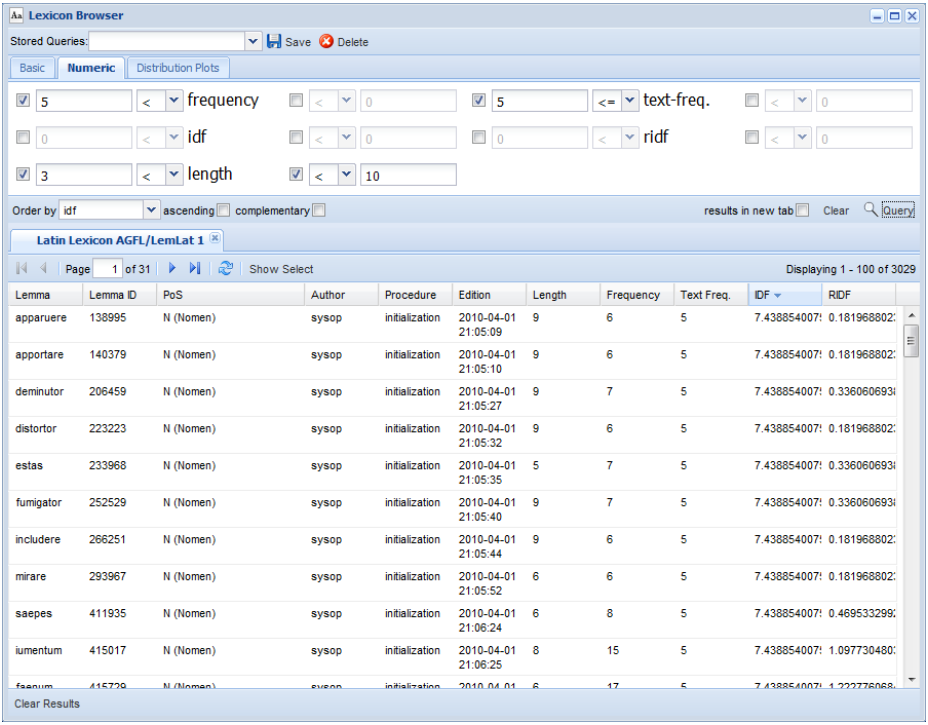


Abbildung 8: Ein Screenshot des *Lexicon Browser* mit Ergebnissen einer Suchanfrage basierend auf einem lateinischen Vollformenlexikon.

Workshops" zu computergestützten Methoden durchgeführt worden, und zwar in Zusammenarbeit mit der Arbeitsgruppe Texttechnologie von Alexander Mehler. In Veranstaltungen dieser drei Lehrformate wurde der eHumanities Desktop dazu genutzt, praktische korpuslinguistische Arbeit in der Geschichtswissenschaft zu leisten.

4.1 Das Format Übung/Seminar

Bei den Übungen und Seminaren mit Studierenden aus dem Grund- und Hauptstudium ging es in erster Linie darum, neue Ansätze zur Erforschung der "Politischen Sprache des Mittelalters" zu vermitteln. Nach einer Einführungsphase zum Konzept des Politischen und seiner Anwendbarkeit auf die Vormoderne oblag es den Studierenden, eigene semantische Untersuchungen zu einzelnen Wörtern zu unternehmen. Dazu wurde ihnen das HSCM-System (siehe Sektion 3.5) zur Verfügung gestellt. Diese Herangehensweise verfolgte zwei Ziele. Zum einen sollte den Studierenden deutlich gemacht werden, dass es

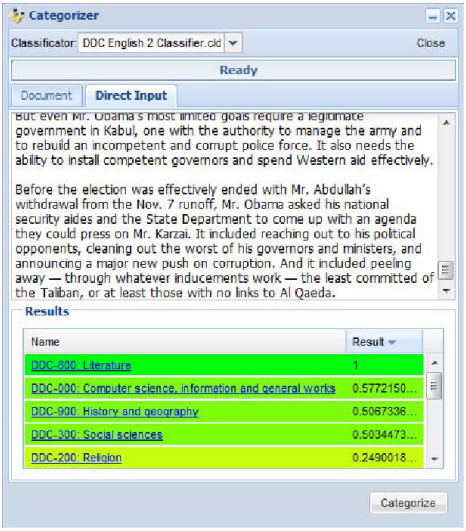


Abbildung 9: Ein Screenshot des *Categorizer* mit dem Ergebnis einer Kategorisierung.

lohnenswert ist, sich von altbekannten Methoden zu verabschieden und in der Forschung unbekanntes Terrain zu betreten. Nicht zuletzt auch, um mit den gewonnen Erkenntnissen bestehende Narrative der Forschung in Frage zu stellen. Zum anderen sollten Kompetenzen im Umgang mit korpuslinguistischen Fragestellungen, Methoden und Werkzeugen eingeübt werden. Dies geschah vor dem Hintergrund eines mediävistischen Erkenntnisinteresses und sollte so den Blick für neue Methoden und Theorien (etwa im Bereich der Historischen Semantik) öffnen.

Die Arbeit mit HSCM hat dabei das eigenständige Forschungsinteresse der Studierenden geweckt. Durch den erleichterten Zugriff auf empirische Daten für die geisteswissenschaftliche Forschung, die HSCM ermöglicht, waren die Teilnehmerinnen und Teilnehmer in der Lage, schnell eigene Ideen zu erproben und zu Ergebnissen zu gelangen. Auf diese Weise lassen sich Studierende besonders gut an Forschungskonzepte und an die Forschungspraxis heranführen.

4.2 Das Format der Studiengruppen

In der Studiengruppe "Historische Semantik"[9] werden gegenüber dem in Sektion 4.1 skizzierten Format der Übung andere Ziele verfolgt. Das Modell für dieses Lehrformat unterscheidet sich wesentlich von dem einer Übung oder auch eines Seminars. Ohne

[9] Dieses Lehrformat wird derzeit an der Goethe-Universität Frankfurt erprobt (`http://www.fzhg.org/front_content.php?idcat=39` [letzter Zugriff 19.04.2011]).

hierarchische Strukturierung dient die Studiengruppe als ein Forum zur Diskussion von forschungsorientierten Problemen und Fragestellungen. Die Gruppe wählt sich ihre Themen selbst aus, trifft sich über mehrere Semester hinweg und verfolgt neben der wissenschaftlichen Erarbeitung eines Themas auch das Ziel der Vernetzung von Wissenschaftlerinnen und Wissenschaftlern an der Goethe-Universität. Sie richtet sich an Professorinnen und Professoren, Mitarbeiterinnen und Mitarbeiter sowie an Studierende. In diesem Zusammenhang wird die Historische Semantik als ein Forschungsprogramm diskutiert. Als Grundlage dieser Diskussionen dienen methodisch-theoretische Basistexte, eigene Arbeiten der Mitglieder, die zum Teil auf dem eHumanities Desktop basieren, sowie Vorträge von eingeladenen Gästen. Außerdem werden Workshops zum Thema der Studiengruppe organisiert.

Auch wenn die Diskussion breit angelegt ist, spielen neue, korpusorientierte und computergestützte Ansätze eine wesentliche Rolle. Durch die verschiedenen Forschungspositionen und Fragestellungen der Mitglieder ergeben sich immer wieder neue interessante Anwendungsbereiche für entsprechende Werkzeuge. In der Studiengruppe wird zudem die basale Erstellung von Korpora diskutiert, und zwar anhand forschungspraktischer Beispiele, um den Teilnehmerinnen und Teilnehmern einen unmittelbaren Einblick in die geisteswissenschaftliche Fachinformatik basierend auf dem eHumanities Desktop zu gewähren.

4.3 Das Workshop-Format

Das dritte Format, in dem die computergestützte Historische Semantik als Lehrinhalt erprobt worden ist, betrifft reine Anwendungsworkshops. Das Ziel dieser Veranstaltungen besteht darin, die Teilnehmerinnen und Teilnehmer mit den Möglichkeiten und Funktionsweisen des eHumanities Desktops und insbesondere des HSCM-Systems vertraut zu machen. Diese Veranstaltungsform adressiert keine spezifische Zielgruppe, sondern richtet sich an Forscherinnen und Forscher, die mit den Werkzeugen arbeiten oder diese diskutieren wollen.

Dieses Veranstaltungsformat erweitert zum einen den Kreis derer, die den Desktop für ihre Forschungen verwenden. Zum anderen liefert es Anregungen zur weiteren Verbesserung und Erweiterung des eHumanities Desktops. Alles in allem dient das Format der Anwendungsworkshops insbesondere dazu, die Vorteile der Programme und Methoden für die geisteswissenschaftliche Forschung und ihre Anwendungsbereiche zu demonstrieren.

5 Diskussion

Nachdem wir drei Einsatzszenarien des eHumanities Desktops in der Lehre umrissen haben, geht es in dieser Sektion um Anforderungen an die Nutzung dieses Systems. Betrachtet man den Kreis der Personen, welche den eHumanities Desktop nutzen, so lassen sich diese erfahrungsgemäß in zwei Gruppen einordnen:

- Auf der einen Seite steht die Zielgruppe der Geisteswissenschaftler, deren Arbeit im Wesentlichen verwendungsorientiert ist. Für diese Nutzergruppe stehen Funktionen aus den Bereichen Suchen, Browsen und Annotieren im Vordergrund. Es geht also um die computerbasierte Repräsentation von Daten, um deren Annotation weit diesseits einer automatischen Weiterverarbeitung. Als Nutzungsprämisse gilt die möglichst niedrige Einstiegsschwelle aller verwendeten Programme: nicht das maschinelle Lernen zur Exploration unbekannter Strukturen, sondern die Datenbankfunktionalität zur Verwaltung von Annotationen und deren Verteilung unter Arbeitsgruppenmitgliedern steht im Fokus.

- Die zweite Zielgruppe bilden die Informatiker, deren Vorgehen modellierungsorientiert ist. Für diese Nutzergruppe stehen Funktionen aus den Bereichen Entwerfen, Modellieren, Entwickeln, Testen und Optimieren im Vordergrund. Vergleichbar mit Systemen wie WEKA (Hall et al., 2009) oder GATE (Bontcheva et al., 2004) erwarten diese Nutzer eine Unterstützung hinsichtlich der Erzeugung automatischer Klassifikatoren oder zumindest die API-basierte Nutzbarkeit von Klassifikatoren von Arbeitsgruppen mit verwandter thematischer Ausrichtung (etwa im Hinblick auf den DDC-Klassifikator des Desktops – siehe Sektion 3.7). Als Nutzungsprämisse gilt die Erweiterbarkeit und prozedurale Vielfalt des Systems, denengegenüber die Datenbankfunktionalität in den Hintergrund rückt.

Dieser Grobeinteilung zweier Nutzergruppen gemäß ergeben sich unterschiedliche Anforderungen an den Ausbau des eHumanities Desktops: auf der einen Seite betrifft dies die Verfügbarkeit von Daten in zunehmender Breite und Tiefe, ob nun auf der Ebene von Korpora oder Lexika. Bei der Analyse von althochdeutschen Texten (Gippert, 2001) tritt beispielsweise die Aufgabe zu Tage, lateinische Lehnwörter zu identifizieren oder Sätze in Althochdeutsch und Latein zu alignieren. Aufgaben dieser Art setzen die Verfügbarkeit von Lexika und Korpora für historische Sprachen voraus, für die vielfach noch immer digitalisierte Ressourcen fehlen. Die hinlängliche Dokumentation aller verwendbaren Werkzeuge wie auch die Kontrollierbarkeit und Rekonstruierbarkeit von Messresultaten stellen weitere typische geisteswissenschaftliche Anforderungen dar. Diese Anforderungen zu erfüllen bedeutet letztlich, vollständige Transparenz und unmittelbare Nachvollziehbarkeit jeder systemseitigen Berechnung zu gewährleisten, ohne auf irgendeine externe Ressource (und sei es eine Literaturreferenz) zurückgreifen zu müssen. In letzter Konsequenz lässt diese Anforderung nur direkte Messungen zu, unter Reduktion des Desktops auf dessen Datenbank- und Retrieval-Funktionalität – unter Ausklammerung jeglicher Form von explorativer Datenanalyse. Für viele geisteswissenschaftliche Anwendungen, auch für weite Teile der historischen Semantik, gilt dies als eine Nutzungsprämisse, so dass es nicht verwundert, dass Module wie das HSCM im Kern Kollokationsanalysen und mengenorientierte Textvergleiche anbieten. Der Grund hierfür besteht darin, dass weitergehende explorative Verfahren hinsichtlich ihrer statistischen Resultate in der geisteswissenschaftlichen Diskussion schwer vermittelbar sind.

An dieser Stelle ist eine Anforderung hervorzuheben, welche die Texttechnologie vor erhebliche Herausforderungen stellt. Dies betrifft jene Art von "pervasive computing", welche sich Geisteswissenschaftler von der Nutzung von Werkzeugen wie dem eHumanities Desktop erwarten. Es geht dabei genauer um die prinzipielle Vernetzung sämtlicher lexikalischer, syntaktischer und textueller Ressourcen, so dass jedes Textsegment durch Aufruf sämtlicher seiner Belegstellen, ob nun in Lexika, Korpora oder Baumbanken, überall im System aktualisiert werden kann: wer beispielsweise in einem lateinischen Korpus wie der PL einen französischen Kommentar findet, will womöglich zu einem französischen Lexikon wechseln können, das die gefundenen Wortformen ausweist, um von dort aus zu einem französischen Korpus wechseln zu können, das weitere Belegstellen derselben Wortformen bereithält u.s.w. Dieses Überall-verfügbar-sein und Weiterverarbeiten-können von Ressourcen, wo immer man sie benötigt, kann als Argument für die Unabdingbarkeit von Anwendungen wie dem eHumanities Desktop gelten, da hier die webbasierte Verfügbarmachung einer objektorientierten Datenbank erst die Voraussetzung für diese Art der Ressourcenvernetzung schafft.

Anforderungen in einem typischerweise geisteswissenschaftlichen Kontext bilden die eine Seite des Anforderungsspektrums. Demgegenüber stehen stärker informatorische Anforderungen wie die prozedurale Erweiterbarkeit des Gesamtsystems und dessen Interoperabilität (Wittenburg et al., 2010). Zusammenfassend gesprochen ergeben sich somit zwei zentrale Forderungen an die Weiterentwicklung des eHumanities Desktops: hinsichtlich des Ausbaus seines prozeduralen Angebots und der Vernetzung von Ressourcen wie auch hinsichtlich der Geschwindigkeit, mit welcher der Desktop weiterentwickelt wird, um immer neuen Anforderungen an eine automatische Verarbeitung geisteswissenschaftlicher Informationsobjekte gerecht zu werden. Denn eines lehrt unsere nunmehr mehrjährige Kooperation von Geisteswissenschaft und Informatik: die Umsetzung hinkt stets der Anforderungsentwicklung hinterher.

6 Zusammenfassung und Schlussfolgerung

Der eHumanities Desktop ist eine webbasierte Schnittstelle für die texttechnologische Arbeit in Forschung und Lehre. Sein Funktionsumfang eignet ihn dazu, eine Vielfalt von Arbeitsfeldern in unterschiedlichen Bereichen der *Digital Humanities* zu bedienen: ausgehend von dem Browsing einzelner Korpora über das Annotieren von Dokumenten und das Bearbeiten einzelner Lexikoneinträge bis hin zur explorativen Analyse von Textvergleichen und dem maschinellen Lernen von Klassifikatoren. Damit adressiert der eHumanities Desktop einen Funktionsumfang, der ihn neben der Forschung auch für den Einsatz in der Lehre eignet. Infolge der webbasierten Architektur des Desktops und seines graphischen Interfaces geht es dabei nicht allein um die Lehre für Computerlinguisten, Texttechnologen oder Informatiker. Vielmehr zeigen die hier beschriebenen Nutzungsszenarien die Einsetzbarkeit des Desktops insbesondere auch für Geisteswissenschaftler. In diesem Zusammenhang haben wir drei Lehrformate diskutiert. Die Erfahrungen, die wir dabei mit dem Lehrinhalt *Historische Semantik* unter Einsatz des *eHumanities Desktops* sammeln konnten, unterscheiden sich je nach Lehrformat. In je-

dem Fall sollte die Anwendungsorientierung im Mittelpunkt jedes dieser Formate stehen. Ein Grund hierfür liegt darin, dass sich theoretische Ansätze und deren methodische Eigenheiten überhaupt erst in der Anwendung als ertragreich erweisen. Dabei bildet die Vermittlung von Fähigkeiten zur Interpretation quantitativer Ergebnisse eine zentrale Lehraufgabe. Erst im Kontext von historischem Grundwissen und wissenschaftlichen Fragestellungen werden die mit Hilfe des eHumanities Desktops ermittelten Zahlen und die damit verknüpften sprachlichen Befunde interpretierbar. Die hier erprobte Interdisziplinarität von Geschichtswissenschaft und Texttechnologie ist offenbar also unabdingbar, um diesen Brückenschlag zwischen qualitativen und quantitativen Methoden zu leisten.

Literatur

Andersen, P. B. (1990). *A Theory of Computer Semiotics: Semiotic Approaches to Construction and Assessment of Computer Systems*. Cambridge University Press, Cambridge.

Bontcheva, K., Tablan, V., Maynard, D., and Cunningham, H. (2004). Evolving GATE to Meet New Challenges in Language Engineering. *Natural Language Engineering*, 10(3/4):349–373.

Evert, S. (2008). Corpora and collocations. In Lüdeling, A. and Kytö, M., editors, *Corpus Linguistics. An International Handbook of the Science of Language and Society*. Mouton de Gruyter, Berlin/New York.

Fodor, J. A. and Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28((1-2)):3–71.

Gippert, J. (2001). TITUS – Alte und neue Perspektiven eines indogermanistischen Thesaurus. *Studia Iranica, Mesopotamica et Anatolica*, 2:46–76.

Gleim, R., Mehler, A., Waltinger, U., and Menke, P. (2009a). eHumanities Desktop – an extensible online system for corpus management and analysis. In *5th Corpus Linguistics Conference, University of Liverpool*.

Gleim, R., Waltinger, U., Ernst, A., Mehler, A., Esch, D., and Feith, T. (2009b). The eHumanities Desktop – an online system for corpus management and analysis in support of computing in the humanities. In *Proceedings of the Demonstrations Session of the 12th Conference of the European Chapter of the Association for Computational Linguistics EACL 2009, 30 March – 3 April, Athens*.

Gleim, R., Warner, P., and Mehler, A. (2010). eHumanities Desktop – an architecture for flexible annotation in iconographic research. In *Proceedings of the 6th International Conference on Web Information Systems and Technologies (WEBIST '10), April 7-10, 2010, Valencia*.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18.

Heyer, G. and Büchler, M. (2010). GI annual meeting 2010. Workshop eHumanities – how does computer science benefit? In Fähnrich, K.-P. and Franczyk, B., editors, *Proceedings of INFORMATIK 2010: Service Science, September 27 – October 01, 2010, Leipzig*, volume 2 of *Lecture Notes in Informatics*. GI.

Heyer, G., Quasthoff, U., and Wittig, T. (2006). *Text Mining: Wissensrohstoff Text*. W3L, Herdecke.

Jussen, B. (2006). "Ordo" zwischen Ideengeschichte und Lexikometrie. Vorarbeiten an einem Hilfsmittel mediävistischer Begriffsgeschichte. In Schneidmüller, B. and Weinfurter, S., editors, *Ordnungskonfigurationen im Hohen Mittelalter*, volume 64 of *Vorträge und Forschungen*, pages 227–256. Thorbecke, Sigmaringen.

Jussen, B., Mehler, A., and Ernst, A. (2007). A corpus management system for historical semantics. *Sprache und Datenverarbeitung. International Journal for Language Data Processing*, 31(1-2):81–89.

Koselleck, R. (1972). Einleitung. In Brunner, O., Conze, W., and Koselleck, R., editors, *Geschichtliche Grundbegriffe. Lexikon zur politisch-sozialen Sprache in Deutschland*, pages V–XXVII. Klett-Cotta, Stuttgart.

Luhmann, N. (1993). *Gesellschaftsstruktur und Semantik. Studien zur Wissenssoziologie der modernen Gesellschaft*, volume 1. Suhrkamp, Frankfurt am Main.

Mehler, A., Diewald, N., Waltinger, U., Gleim, R., Esch, D., Job, B., Küchelmann, T., Pustylnikov, O., and Blanchard, P. (2011). Evolution of Romance language in written communication: Network analysis of late Latin and early Romance corpora. *Leonardo*, 44(3).

Mehler, A., Gleim, R., Ernst, A., and Waltinger, U. (2008). WikiDB: Building interoperable wiki-based knowledge resources for semantic databases. *Sprache und Datenverarbeitung. International Journal for Language Data Processing*, 32(1):47–70.

Mehler, A., Gleim, R., Waltinger, U., and Diewald, N. (2010). Time series of linguistic networks by example of the Patrologia Latina. In Fähnrich, K.-P. and Franczyk, B., editors, *Proceedings of INFORMATIK 2010: Service Science, September 27 – October 01, 2010, Leipzig*, volume 2 of *Lecture Notes in Informatics*, pages 609–616. GI.

Mehler, A., Gleim, R., Waltinger, U., Ernst, A., Esch, D., and Feith, T. (2009). eHumanities Desktop – eine webbasierte Arbeitsumgebung für die geisteswissenschaftliche Fachinformatik. In *Proceedings of the Symposium "Sprachtechnologie und eHumanities", 26.–27. Februar, Duisburg-Essen University.*

Migne, J.-P., editor (1844–1855). *Patrologiae cursus completus: Series latina*, volume 1–221. Chadwyck-Healey, Cambridge.

Petitot, J., Varela, F. J., Pachoud, B., and Roy, J.-M. (2000). *Beyond the Gap: An Introduction to Naturalizing Phenomenology.* Stanford University Press, Stanford.

Radl, C. (2007). Die digitalen Monumenta Germaniae Historica. In historischer Forschungseinrichtungen in der Bundesrepublik Deutschland e.V., A., editor, *Jahrbuch der historischen Forschung in der Bundesrepublik Deutschland: Berichtsjahr 2006*, pages 69–74. München.

Rexroth, F. (2009). Politische Rituale und die Sprache des Politischen in der historischen Mittelalterforschung. In DeBenedictis, A., Corni, G., Mazohlm, B., and Schorn-Schütte, L., editors, *Die Sprache des Politischen in actu. Zum Verhältnis von politischem Handeln und politischer Sprache von der Antike bis ins 20. Jahrhundert*, pages 71–90. Vandenhoeck & Ruprecht, Göttingen.

Rieger, B. (1998). Warum fuzzy Linguistik? Überlegungen und Ansätze zu einer computerlinguistischen Neuorientierung. In Krallmann, D. and Schmitz, H. W., editors, *Perspektiven einer Kommunikationswissenschaft. Internationales Gerold Ungeheuer Symposium, Essen 1995*, pages 153–183. Nodus, Münster.

Rieger, B. B. (1989). *Unscharfe Semantik: Die empirische Analyse, quantitative Beschreibung, formale Repräsentation und prozedurale Modellierung vager Wortbedeutungen in Texten.* Peter Lang, Frankfurt a. M.

Rieger, B. B. (2003). Semiotic cognitive information processing: Learning to understand discourse. A systemic model of meaning constitution. In Kühn, R., Menzel, R., Menzel, W., Ratsch, U., Richter, M. M., and Stamatescu, I. O., editors, *Adaptivity and Learning. An Interdisciplinary Debate*, pages 347–403. Springer, Berlin.

Skinner, Q. (2003). *Visions of Politics (reprint)*, volume 1: Regarding Method. Cambridge University Press, Cambridge.

Steinmetz, W. (2008). Vierzig Jahre Begriffsgeschichte. The State of the Art. In Kämper, H. and Eichinger, L. M., editors, *Sprache – Kognition – Kultur. Sprache zwischen mentaler Struktur und kultureller Prägung [Vorträge der Jahrestagung 2007 des Instituts für Deutsche Sprache]*, pages 174–197. De Gruyter, Berlin.

TEI Consortium, editor (2010). *TEI P5: Guidelines for Electronic Text Encoding and Interchange.* TEI Consortium, http://www.tei-c.org/Guidelines/P5/.

von Ahn, L. (2008). Human computation. In *IEEE 24th International Conference on Data Engineering (ICDE 2008)*, pages 1–2. IEEE.

Wagner, B., Mehler, A., Wolff, C., and Dotzler, B. (2009). Bausteine eines Literary Memory Information System (LiMeS) am Beispiel der Kafka-Forschung. In *Proceedings of the Symposium "Sprachtechnologie und eHumanities", 26.–27. Februar, Duisburg-Essen University.*

Waltinger, U. (2010). *On Social Semantics in Information Retrieval.* Phd thesis, Bielfeld University, Germany.

Waltinger, U., Mehler, A., Lösch, M., and Horstmann, W. (2011). Hierarchical classification of OAI metadata using the ddc taxonomy. In Bernardi, R., Chambers, S., Gottfried, B., Segond, F., and Zaihrayeu, I., editors, *Advanced Language Technologies for Digital Libraries (ALT4DL)*, LNCS. Springer-Verlag, Berlin.

Wittenburg, P., Hinrichs, E. W., Broeder, D., and Zastrow, T. (2010). eHumanities – können wir die Komplexität beherrschen? In Fähnrich, K.-P. and Franczyk, B., editors, *Proceedings of INFORMATIK 2010: Service Science, September 27 – October 01, 2010, Leipzig*, volume 2 of *Lecture Notes in Informatics*, pages 530–541. GI.

Michael Beißwenger & Angelika Storrer

Digitale Sprachressourcen in Lehramtsstudiengängen: Kompetenzen – Erfahrungen – Desiderate

1 Einleitung

In diesem Artikel beschreiben wir,

- wie sich die Arbeit mit digitalen Sprachressourcen in die Curricula der Hochschulgermanistik und der Lehrpläne für das Fach Deutsch an Schulen integrieren lässt und weshalb wir die Vermittlung entsprechender Kompetenzen in Lehramtsstudiengängen für wichtig und hochgradig berufsfeldrelevant halten,
- wie wir an der Technischen Universität Dortmund in den Bereichen Linguistik und Sprachdidaktik Sprachressourcen in der Lehre einsetzen (und welche),
- welche Erfahrungen wir dabei gemacht haben und welche Wünsche und Anregungen zur Erleichterung des didaktischen Einsatzes von Sprachressourcen sich daraus ableiten lassen.

Die allgemeinen Überlegungen zur Relevanz des Themas für die Lehrerbildung und zur curricularen Einbettung in entsprechende Studiengänge werden ergänzt um die Beschreibung zweier Seminarkonzepte zu den Themen „Korpusgestützte Sprachanalyse“ und „Internetbasierte Kommunikation“ (Abschnitte 3 und 4). Im einen Fall ist die Vermittlung von Methodenkompetenz in Bezug auf die Nutzung digitaler Sprachressourcen selbst Gegenstand der Veranstaltung, im anderen Fall werden Sprachressourcen punktuell als Hilfsmittel für die Durchführung kleiner Analyseprojekte genutzt. Beide Seminarkonzepte wurden bereits mehrfach erprobt und dabei z.T. auch weiterentwickelt. In Abschnitt 5 formulieren wir auf der Grundlage unserer bisher gemachten Erfahrungen Desiderate für die Ausgestaltung digitaler Sprachressourcen in diesem Bereich.

Sämtliche im Artikel mit Kurztiteln erwähnten Sprachressourcen sind in Abschnitt 6.2 mit ihren vollständigen Bezeichnungen und zugehörigen URL-Verweisen aufgeführt.

2 Motivation und curriculare Einbettung

2.1 Generelle Überlegungen

Zwei generelle Überzeugungen motivieren den Einsatz von digitalen Sprachressourcen in der Ausbildung von Lehramtsstudierenden am Institut für deutsche Sprache und Literatur der TU Dortmund:

(1) Die fachwissenschaftliche Ausbildung in deutscher Linguistik sollte nicht nur Theorien, Beschreibungskategorien und -modelle vermitteln, sondern auch die Methoden, die der Modell- und Kategorienbildung zugrunde liegen. Erst dies befähigt angehende LehrerInnen dazu, didaktische Konzepte des „entdeckenden Lernens“, wie sie z.B. in der „Grammatikwerkstatt“ (EISENBERG & MENZEL 1999) oder im funktionalen Grammatikunterricht (HOFFMANN 2006) angestrebt werden, im schulischen Sprachunterricht souverän umzusetzen. Da

digitale Sprachressourcen, insbesondere digitale Text- und Gesprächskorpora, für die empirisch geleitete linguistische Kategorien- und Modellbildung an Bedeutung gewinnen, werden entsprechende Methoden auch für Lehramtsstudierende relevant, selbst wenn diese an eigener linguistischer Forschung nicht primär interessiert sind. Weiterhin sind digitale Sprachressourcen im Internet natürlich auch eine Fundgrube für Unterrichtsmaterial im Rahmen von internetbasierten Didaktikkonzepten wie z.B. WebQuests.

(2) Die Kompetenz zur kritischen Bewertung und aufgeklärten Nutzung digitaler Sprachressourcen zur deutschen Sprache (lexikalische und enzyklopädische Ressourcen, Korpora) wird für angehende DeutschlehrerInnen zunehmend relevant. Gerade weil viele verschiedene Typen von Ressourcen im Internet verfügbar sind und sich als potenzielle Datenquellen zur deutschen Sprache anbieten (vgl. ENGELBERG & LEMNITZER 2009; STORRER 2010), ist es wichtig, Schülerinnen und Schülern Leitlinien für die Bewertung der Qualität und der Zuverlässigkeit dieser Daten zu vermitteln. Dies gilt für kollaborativ aufgebaute lexikalische Ressourcen, z.B. *Wiktionary*, das mehrsprachige *Dict.cc*, den *OpenThesaurus*, ebenso wie für retrospektiv digitalisierte Wörterbücher, z.B. *DWDS-WDG* oder *DWB-Online*, deren Quellenwert an den ursprünglichen Entstehungskontext angebunden werden muss. Auch für die aufgeklärte und kundige Nutzung digitaler Sprachkorpora müssen Kriterien vermittelt werden, anhand derer Studierende bewerten können, welche Aussagekraft die bei der Korpusrecherche erzielten Ergebnisse für eine bestimmte Fragestellung haben.

Wichtig ist uns, dass wir die Nutzung von digitalen Sprachressourcen nicht als eine linguistische „Schule“ einführen, sondern als einen methodischen Zugang zu Sprache, der parallel und ggf. auch in Kombination mit anderen methodischen Zugängen dazu beiträgt, unser Wissen über sprachliche Strukturen und Funktionen zu erweitern. Gerade damit der Stellenwert datengestützter Zugänge im methodischen Spektrum der Sprachwissenschaften besser bestimmt werden kann, muss nach unserer Erfahrung nicht nur das computertechnische Rüstzeug – z.B. die Syntax von Abfragesprachen – vermittelt werden, sondern auch die Kompetenz zur kritischen Bewertung der Qualität von Daten und der Aussagekraft von datengestützten Analysen. Wir ermuntern die Studierenden dazu, sich vor der Nutzung von digitalen Ressourcen Fragen wie die folgenden zu beantworten:

- Wer baut die Ressource auf / bearbeitet sie redaktionell / bietet sie an?
- Welche Maßnahmen der Qualitätskontrolle werden von Seiten der Anbieter thematisiert? Wie kann man als Nutzer die Qualität der Daten evaluieren? Welche qualitativen Aspekte sind für meine Fragestellung besonders relevant.
- Welchen Sprachausschnitt beschreibt/dokumentiert die Ressource? Wie umfangreich ist sie (im Vergleich zu Ressourcen, die ggf. als Alternativen zur Verfügung stehen)?
- Eignet sich die Ressource überhaupt für die Bearbeitung meiner Fragestellungen; welche Einschränkungen gibt es?
- Wie ist die Ressource strukturiert? Welche Typen von Daten sind systematisch mit welchen Beschreibungskategorien erfasst?

Der Umgang mit digitalen Textkorpora stellt dabei besonders hohe Ansprüche: Für die vergleichende Bewertung von Korpusressourcen müssen grundlegende Konzepte der Korpuslinguistik (Annotation, Metadaten, Lemmatisierung, Wortartentagging etc.) und der Qualitätsbewertung verstanden worden sein. Wie wir in Abschnitt 3 noch weiter ausführen

werden, ist es nach unserer Erfahrung sinnvoll, die Vermittlung dieser Konzepte mit kleinen praktischen Aufgaben und Beispielen zu verknüpfen und auf der Basis der dabei gemachten Erfahrungen Qualitätsaspekte und methodische Herausforderungen zu reflektieren.

Wie und mit welchen Kompetenzzielen digitale Ressourcen im Unterricht in verschiedenen Schulstufen nutzbar sind, muss noch weiter erprobt werden. Nach unserer Erfahrung sind Studierende sehr gut motivierbar, didaktische Konzepte zu entwickeln und in schulischen Praxisphasen zu erproben; sie stoßen dabei erfahrungsgemäß auf großes Interesse bei den Schülerinnen und Schülern aller Altersstufen und meist auch auf Interesse seitens der Schulleitungen und der betreuenden Lehrerinnen und Lehrer. Die Praxisphasen in den Lehramtsstudiengängen bieten hier Anknüpfungspunkte für Experimente.

2.2 Anknüpfungspunkte an schulische Lehrpläne

Bislang ist der Einsatz digitaler Sprachressourcen nicht explizit in den Kompetenzprofilen schulischer Lehrpläne verankert. Die verschiedenen Formen der Ressourcennutzung und die dafür zu vermittelnden Kompetenzen sind aber problemlos an übergreifende Kompetenzentwicklungsziele anschließbar. Im Folgenden seien hierfür nur einige Beispiele aus den Lehrplänen des Landes Nordrhein-Westfalen genannt:

(1) *Wissenschaftspropädeutik*: Als eine der beiden zentralen Aufgaben der gymnasialen Oberstufe nennen die Richtlinien und Lehrpläne des Landes Nordrhein-Westfalen die *Wissenschaftspropädeutik* – im Sinne eines „wissenschaftsorientierte[n] Lernen[s], das durch Systematisierung, Methodenbewusstsein, Problematisierung und Distanz gekennzeichnet ist“ (RLP SEK II GYGE DEUTSCH NRW: XII). Ein wichtiges Vermittlungsziel im Rahmen des wissenschaftspropädeutischen Lernens besteht darin, die Schülerinnen und Schüler dazu zu befähigen, „grundlegende wissenschaftliche Erkenntnis- und Verfahrensweisen systematisch [zu] erarbeiten“ und dabei zu lernen, „Aufgabenstellung[en] selbstständig zu strukturieren, die erforderlichen Arbeitsmethoden problemangemessen und zeitökonomisch auszuführen, Hypothesen zu bilden und zu prüfen und die Arbeitsergebnisse angemessen darzustellen“ (ebd.). Die Verknüpfung unterrichtlicher Sprachreflexion mit einer Analyse authentischer Sprachdaten (Korpora) kann in diesem Zusammenhang einerseits Zugänge zur Arbeitsweise einer modernen datengestützten Linguistik eröffnen und andererseits einen Ausgangspunkt für die Vermittlung allgemein-wissenschaftspropädeutischer Grundlagen zum Verhältnis von wissenschaftlicher Theoriebildung und Empirie im Allgemeinen bilden. Da die Wissenschaftspropädeutik eine Aufgabe nicht nur des Deutsch-, sondern jedweden Fachunterrichts in der gymnasialen Oberstufe darstellt, sind hier auch fächerübergreifende Konzepte denkbar.

(2) *Methodenkompetenz*: In der didaktischen Konzeption des Faches Deutsch wird neben der sprachlichen, kulturellen, ästhetischen und ethischen Kompetenz die *Methodenkompetenz* als eines von fünf zentralen Vermittlungszielen herausgestellt. Der Fokus liegt dabei auf „Methoden sprachlichen Arbeitens (Methoden des Verstehens, Methoden sprachanalytischer Arbeit, Methoden schriftlicher Darstellung, Methoden mündlicher Verständigung, Methoden produktionsorientierten Arbeitens), die gleichzeitig Unterrichtsgegenstand, fachliche Verfahrensweisen und Lernstrategien darstellen“ (RLP SEK II GYGE DEUTSCH NRW: 6). In allen Teilaspekten liegt die Arbeit mit digitalen Sprachressourcen nahe, z.B.:

- Als Methoden für das Verstehen und die Erschließung von Texten bilden Kenntnisse im Umgang mit Nachschlagewerken und zu den in Nachschlagewerken unterschiedlichen Typs verfügbaren Informationstypen eine unerlässliche Grundlage. Die „Beschaffung und Bearbeitung von Informationen“ und die „Nutzung von Hilfsmitteln, auch in elektronischen Netzen“ (ebd.: 27) werden in diesem Zusammenhang als zentrale Vermittlungsziele im Bereich „Methoden des Textverstehens“ benannt – ein Bereich, in den sich die Arbeit mit digitalen Sprachressourcen (z.B. digitalen Allgemeinwörterbüchern zur deutschen Gegenwartssprache, online verfügbaren Autoren- und historischen Wörterbüchern) unmittelbar integrieren lässt.
- Zu den „Methoden schriftlicher Darstellung“ zählt sicher auch das Nachschlagen in Online-Nachschlagewerken zur deutschen Sprache; hierfür müssen Kompetenzen zur Bewertung von Qualität und Zuverlässigkeit vermittelt werden. Eine „effektive Nutzung der fachspezifischen Informations- und Kommunikationsangebote in Bibliotheken und elektronischen Netzen“ (ebd.: 28) ist hier explizit als ein Kompetenzziel benannt, gerade auch im Zusammenhang mit der wissenschaftspropädeutischen Aufgabe des Deutschunterrichts.
- Für die Vermittlung von „Methoden sprachanalytischer Arbeit“ bietet sich, insbesondere in Kombination mit dem Bereich „Reflexion über Sprache“, die Arbeit mit digitalen Korpora an. Mit Blick auf die Aufgabe des sprachbezogenen Deutschunterrichts, die Schülerinnen und Schüler zu einem „norm- und regelbewussten Sprechen und Schreiben“ und zu einer funktional angemessenen Differenzierung von Sprachverwendungsweisen und stilistischen Varianten zu befähigen (ebd.: 22), kann die Arbeit mit Korpora eine wichtige Bereicherung des didaktischen Instrumentariums darstellen, weil sie bei entsprechender Einführung und der Formulierung geeigneter Arbeitsaufgaben dazu genutzt werden kann, die Schülerinnen und Schüler auf die Besonderheiten stilistischer Variation in unterschiedlichen Zweckbereichen und situativen Kontexten sprachlichen Handelns selbst aufmerksam werden zu lassen.

Um den Umgang mit digitalen Sprachressourcen in der Schule vermitteln zu können, müssen Lehrerinnen und Lehrer selbst reflektierte Kenntnisse zu entsprechenden Ressourcen besitzen und über Möglichkeiten und Rahmenbedingungen orientiert sein, diese didaktisch sinnvoll im Unterricht einzusetzen. Im folgenden Abschnitt möchten wir deshalb am Beispiel der Dortmunder Lehramtsstudiengänge skizzieren, wie sich solche Kompetenzen in die Curricula integrieren lassen.

2.3 Anknüpfungspunkte an universitäre Lehramtsstudiengänge

Zu den übergreifenden Zielen der Dortmunder Lehramtsstudiengänge im Fach Deutsch zählt die Vermittlung von „Erfahrungen in der selbständigen wissenschaftlichen Arbeit“ (FSB GERM BFP DORTMUND: 39). Der Umgang mit digitalen Sprachressourcen kann hierzu in mehrerlei Hinsicht beitragen: Digitale Ressourcen wie Korpora und Informationssysteme zur Lexik und Grammatik können für die eigenständige wissenschaftliche Arbeit genutzt werden; hierbei werden deren Potenziale und Besonderheiten im Rahmen konkreter Nutzungsszenarien (z.B. Nachschlageaufgabe, Analyseprojekt) deutlich – eine wichtige Voraussetzung, um einschätzen zu können, ob und wie diese Ressourcen auch im schulischen Deutschunterricht einsetzbar sind. In den fachdidaktischen Anteilen des Studiums kann der

Einsatz digitaler Sprachressourcen in der Schule explizit thematisiert und im Rahmen von Praxisphasen erprobt werden. Dieser curriculare Aufbau – Einführung in den Umgang mit Sprachressourcen → Anwendung in fachwissenschaftlichen Analyseprojekten → Entwicklung fachdidaktischer Konzepte für den schulischen Einsatz → Erprobung und Evaluation in der Schulpraxis – ist prinzipiell in alle Phasen und Aspekte des Lehramtsstudiums (die einführenden wie die aufbauenden, die fachwissenschaftlichen wie fachdidaktischen) integrierbar.

In den Lehramtsstudiengängen an der TU Dortmund werden bereits in den sprachwissenschaftlichen Pflichtveranstaltungen (eine 4-stündige „Einführung in die Sprachwissenschaft“ und ein 2-stündiges Proseminar „Grundlagen der Grammatik“) digitale Sprachressourcen – insbesondere *Canoo.net*, *ProGr@mm* und *Grammis* – für kleine Recherche- und Analyseaufgaben genutzt; in diesem Zusammenhang werden auch Aspekte der Qualität und Verlässlichkeit von digitalen Quellen thematisiert.

In weiterführenden Wahlpflichtveranstaltungen werden diese Grundkenntnisse vertieft und um weitere Ressourcentypen – z.B. digitale Korpora und lexikalische Informationssysteme – erweitert. Dabei lassen sich zwei Typen der Integration digitaler Sprachressourcen in die didaktischen Konzepte der Lehrveranstaltungen unterscheiden:

Typ 1: Seminare, in denen digitale Sprachressourcen den *Gegenstand* der Veranstaltung bilden (z.B. zu den Themen „Digitale Nachschlagewerke zur deutschen Sprache“, „Korpusgestützte Sprachanalyse“, „Internetbasierte Lexikographie“). In Veranstaltungen dieses Typs werden – je nach Seminarthema – verschiedene Kombinationen von digitalen Korpora, lexikalischen und grammatischen Informationssystemen und digitalen Wörterbüchern genutzt. Lehr-/Lernziele sind die Vermittlung von Methodenwissen zum Umgang mit Sprachressourcen, die Erarbeitung von Qualitätskriterien und die kritische Reflexion ihres Nutzungspotenzials in verschiedenen Anwendungsfeldern (insbesondere Sprachdidaktik und Lexikographie). In Abschnitt 3 wird ein Seminarkonzept dieses Typs im Detail vorgestellt.

Typ 2: Seminare, in denen digitale Sprachressourcen als *Hilfsmittel und empirische Basis* für Analyseprojekte der Studierenden genutzt werden (z.B. zu den Themen „Wortbildung und Wortschatzentwicklung“, „Phraseologie des Deutschen“, „Orthographie des Deutschen“, „Mediale Bedingungen des kommunikativen Handelns“, „Internetbasierte Kommunikation“). Die Art der Analyseaufgaben und die dafür geeigneten Ressourcen ergeben sich aus dem jeweiligen Seminarthema. Zum Einsatz kamen bislang z.B. die Korpora und Wörterbücher im *DWDS*-Portal, die *Canoo*-Spezialwörterbücher zur Wortbildung und zur Orthographie, die *Wortwarte* und das *Szenesprachenwiki* zur Wortschatzentwicklung im Gegenwartsdeutschen, das Neologismen- und das phraseologische Wörterbuch in der *eLexiko*-Komponente von *OWID* sowie Spezialkorpora zur internetbasierten Kommunikation. In Abschnitt 4 werden wir ein Seminarkonzept dieses Typs zum Thema Internetbasierte Kommunikation vorstellen, das mit am Lehrstuhl aufgebauten Korpora zur Chat-Kommunikation arbeitet.

3 Seminar „Korpusgestützte Sprachanalyse“

3.1 Gegenstand, Szenario, genutzte Ressourcen

Für die kundige Nutzung digitaler Textkorpora benötigt man sowohl korpuslinguistisches Grundwissen als auch Leitlinien für die methodische Planung und Durchführung korpusgestützter Analysen. An der TU Dortmund haben wir ein Seminarkonzept erprobt und weiterentwickelt, um diese Kombination von korpuslinguistischen und empirisch-methodischen Kompetenzen aufzubauen. Die Seminare sind verankert im Wahlpflichtbereich des Moduls „Anwendungsfelder der Sprach-, Literatur- und Medienwissenschaft“ der Dortmunder Lehramtsstudiengänge. Das Seminarkonzept umfasst einen Teil, in dem die methodischen Grundlagen der empirischen Sprachanalyse vermittelt werden, z.B. die Unterscheidung von hypothesenerkundendem und hypothesenprüfendem Vorgehen, Leitlinien zur Formulierung falsifizierbarer Hypothesen und zur Bewertung der Validität und Reliabilität von Ergebnissen, die Frage der Repräsentativität von Stichproben, der Stellenwert von Signifikanztests etc. Als Basisliteratur dienen BORTZ & DÖRING (2006), für den Umgang mit Excel zusätzlich ALBERT & KOSTER (2002).[1] Ein weiterer Bestandteil des Seminarkonzepts ist ein Überblick über digitale Sprachressourcen mit dem Fokus auf deutschen Korpora (vgl. die Auswahl in STORRER 2011: Kap.3). Im Anschluss an diesen Überblick werden Grundbegriffe der Korpuslinguistik eingeführt, z.B. die Unterscheidung von Primär- und Metadaten, Formen der linguistischen Aufbereitung wie Lemmatisierung oder POS-Tagging, Verfahren der automatischen Kollokationsanalyse, die Qualitätsbewertung von Suchanfragen durch Genauigkeit (*precision*) und Ausbeute (*recall*) etc. Basisliteratur für diesen Teil bildet die Einführung von LEMNITZER & ZINSMEISTER (2006), ergänzt um MCENERY et al. (2006). Erstmalig wurde das Seminar im WS 2006/07 angeboten. Dabei handelte es sich um eine wöchentlich stattfindende Veranstaltung, in der zunächst die Vermittlung der o.g. Konzepte und Methoden am Beispiel von Korpusarbeiten anderer ForscherInnen im Vordergrund stand. Erst gegen Ende des Seminars konzipierten die Studierenden eigene Fragestellungen, die sie nach Seminarende mit Hilfe von Korpusanalysen im Rahmen von Hausarbeiten weiterverfolgten. Der dabei entstandene Betreuungsbedarf sowie entsprechende Rückmeldungen zeigten, dass den Studierenden die Relevanz der im Seminar besprochenen korpuslinguistischen Grundbegriffe und der Nutzwert der linguistischen Aufbereitung (Lemmatisierung, POS-Tagging) erst später beim eigenen Arbeiten deutlich geworden waren. Insbesondere die Möglichkeiten, durch geschickte Abfragestrategien die Ergebnisse von Korpusanalysen zu optimieren, wurden erst im Zuge der eigenen Projekte als sinnvoll und wichtig entdeckt (obwohl sie im Seminar an Beispielen behandelt worden waren).

[1] Wir beziehen uns hier auf die Auflagen, die bei der ersten Durchführung im Wintersemester 2006/07 verfügbar waren; in späteren Seminaren wurden jeweils die aktuellsten Auflagen genutzt.

1) In Bodo Mrotzeks „Lexikon der bedrohten Wörter“ sind folgende Wörter verzeichnet: *Dusel, Pappenheimer, Schabernack, Schelm.*
Bestätigen die Daten des DWDS-Kernkorpus und des DWDS-Zeitungskorpus „Die ZEIT“ die Annahme, dass diese Wörter vom „Aussterben“ bedroht sind?
2) Versuchen Sie, für das Verb „eintrudeln“ möglichst viele Belege zu finden, d.h. auch Belege, in denen die Partikel mit dem Verb die Satzklammer bildet (…*trudelt* … *ein*).
3) Im Zuge der Orthographiereform wurde der Bereich der Getrennt- und Zusammenschreibung kontrovers diskutiert. Untersuchen Sie im DWDS-Kernkorpus, welche der Schreibalternativen von „Eis laufen“ vs. „eislaufen“ in welchen Zeitabschnitten des 20. Jahrhunderts häufiger belegt ist.

Notieren Sie bitte bei der Lösung der Aufgaben 1–3 nicht nur die Ergebnisse, sondern auch die Abfragen, die Sie gemacht haben und ggf. Auffälligkeiten in den dabei erzielten Trefferlisten.

4) Es gilt als ein Trend der (gesprochenen) Gegenwartssprache, dass „weil“ nicht mehr nur als unterordnende Konjunktion (mit Verbendstellung: … *weil das Wetter schlecht ist*), sondern auch als nebenordnende Konjunktion (mit Verbzweitstellung … *weil das Wetter ist schlecht*) verwendet wird.

Mit der Suchanfrage **"weil \, "** kann man im Korpus gesprochener Sprache des DWDS die Suche nach „falschen“ weil-Sätzen auf 117 eingrenzen.

- Welche Arten von Pseudotreffern gibt es?
- Inwiefern könnten die Daten belegen, dass es sich tatsächlich um eine neuere Entwicklung handelt?
- Kann man die Präzision der Abfrage noch weiter optimieren?
- Mit welchen Strategien kann man die Vollständigkeit der Abfrage noch weiter erhöhen?

Abbildung 1: Beispiele für Rechercheaufgaben in den Praxisteilen des Seminars „Korpusgestützte Sprachanalyse“.

Aus dieser Erfahrung heraus wurde das Konzept in den Folgeseminaren (im SS 2008 und im SS 2009) so umgestaltet, dass die Vermittlung methodischer und konzeptioneller Grundlagen jeweils verzahnt war mit konkreten kleinen Analyseaufgaben, an denen die Relevanz der Konzepte und Lösungsstrategien für bestimmte Problemstellungen deutlich wird. Dabei erwies sich die Organisationsform als einwöchiges Blockseminar als günstig, das in einem Seminarraum mit angeschlossenem Rechnerraum (mit 30 Plätzen) durchgeführt wurde. Die in den Vormittagssitzungen vermittelten Methoden, Konzepte und Abfragestrategien wurden von Arbeitsgruppen nachmittags am Beispiel kleiner Analyseaufgaben erprobt, dabei standen AnsprechpartnerInnen (Seminarleiterin/TutorInnen) für Rückfragen direkt zur Verfügung. Die Erfahrungen und Ergebnisse der Arbeitsgruppen wurden wiederum im Plenum vorgestellt und diskutiert; diese Diskussionen bildeten gute Anknüpfungspunkte für die zunächst anwendungsunabhängig eingeführten Konzepte und Methoden. Die Studierenden waren für die praktischen Arbeiten sehr gut zu motivieren und entwickelten viel Kreativität beim Experimentieren und Optimieren von Abfragen. Um Unterschiede im Tempo der Arbeitsgruppen aufzufangen, gab es in den AG-Phasen immer mehrere Aufgaben, die in frei wählbarer Reihenfolge, aber nicht unbedingt vollständig zu bearbeiten waren. Um zu vermeiden, dass sich die Studierenden in zu viele unterschiedliche Abfragesprachen eindenken

mussten, bezogen sich die meisten Aufgaben auf die Online-Schnittstelle zu den *DWDS*-Korpora (GEYKEN 2007). Abb. 1 zeigt einige Beispiele für solche Aufgaben; weitere Beispiele und Lösungskommentare dazu – mit Fokus auf dem Anwendungsbereich Lexikographie – finden sich in den Übungen zu STORRER (2011). Am Ende des Blockseminars wurden Projektpläne für korpusbasierte Studien entworfen und im Plenum diskutiert. Als Leitfaden für die Projektplanung diente das Schema in Abb. 2, das vorab im Seminar an Beispielen aus eigenen korpusgestützten Arbeiten bzw. von der Seminarleiterin betreuten Staatsexamens- und Masterarbeiten vorgestellt und konkretisiert worden war. Die Studierenden konnten entweder vorgegebene Fragestellungen bearbeiten oder Pläne für eigene Projekte entwerfen. Die meisten Studierenden entschieden sich für Letzteres und einige realisierten die geplanten Studien im Rahmen von Hausarbeiten (sofern eine Hausarbeit für die von den Studierenden gewünschte Kreditierung erforderlich war). Der Betreuungsaufwand bei diesen Hausarbeiten erwies sich als erheblich geringer als beim ersten, weniger praxisorientierten Seminardurchlauf, da durch die Praxisphasen die Kompetenz zur Formulierung geeigneter Abfragen schon vorhanden war und der Aufwand für die intellektuelle Nachanalyse besser abgeschätzt werden konnte.

Die Fragestellungen, die in den Praxisteilen bearbeitet werden, sind wegen der begrenzten Zeit, die für die Analysen zur Verfügung steht, natürlich vergleichsweise simpel. Deshalb wurde das Spektrum der Möglichkeiten korpusunterstützten und korpusbasierten Arbeitens erweitert durch Hinweise auf komplexere Studien in der weiterführenden Literatur, z.B. auf die Spezialartikel der HSK-Bände zur Korpuslinguistik (LÜDELING & KYTÖ 2008/2009) und zu korpusbasierten Ansätzen in speziellen linguistischen Phänomen- und Anwendungsbereichen (z.B. Phraseologie, Wortbildung; Lexikographie, Grammatikographie). Allerdings sprengen die meisten publizierten Studien den zeitlichen Rahmen, der für studentische Arbeiten und erst recht später für kleine Korpusstudien im schulischen Deutschunterricht zur Verfügung steht. Für das Anliegen, das Arbeiten mit Korpora auch für die Lehrerbildung attraktiv zu machen, ist es deshalb hilfreich, Studierende über kleine Übungen an die Chancen, aber auch an die Grenzen der aktuellen Korpustechnologie heranzuführen. Gerade computerlinguistisch nicht ausgebildete Korpusnutzer überschätzen nämlich tendenziell die Möglichkeiten der automatischen Korpusanalyse und unterschätzen den manuell-intellektuellen Aufwand, der für die Beantwortung vieler linguistischer Fragestellungen immer noch betrieben werden muss. Dies liegt daran, dass sie die automatischen Verfahren zur (computer-)linguistischen Aufbereitung von Korpora nicht durchschauen und deshalb nicht einschätzen können, in welchen Bereichen beim aktuellen Stand der Technologie noch mit vielen Fehlern gerechnet werden muss und welche (prinzipiell interessanten) Fragen ggf. auch überhaupt noch nicht mit vertretbarem Aufwand in digitalen Korpora untersucht werden können. Wenn die Kompetenzen zur Auswahl geeigneter Fragestellungen und geeigneter Heuristiken für die Datenerhebung und -analyse fehlen, kann das Arbeiten mit Korpora schnell als mühselige „Erbsenzähl-Linguistik“ erscheinen. Es ist deshalb ein wichtiges Anliegen des Seminarkonzepts, durch die kleinen Analyseaufgaben typische Fehlerquellen und Grenzen der aktuellen Korpustechnologie deutlich zu machen.

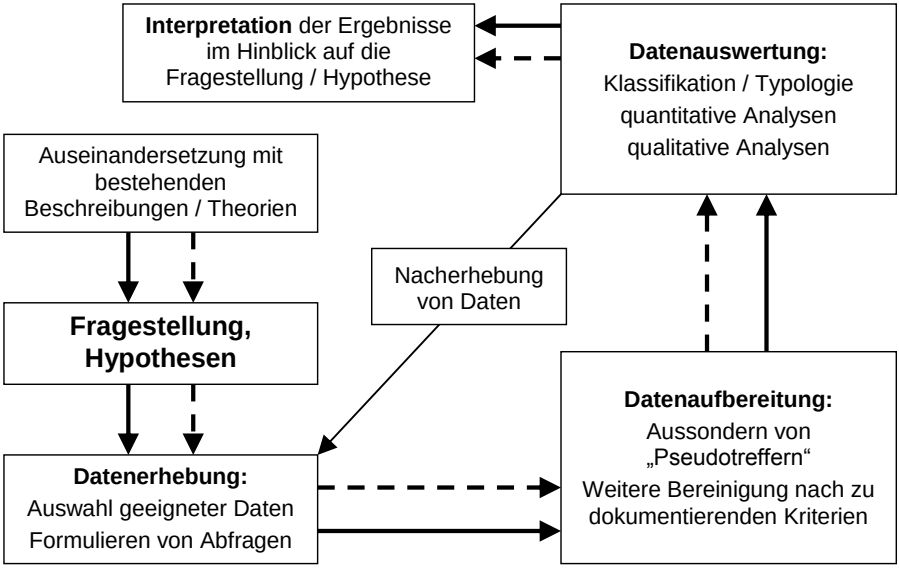


Abbildung 2: Schema für die Planung korpusbasierter Studien.

3.2 Erfahrungen mit Fehlerquellen und Grenzen der aktuellen Korpustechnologie

Im Folgenden möchten wir einige Beispiele für solche Fehlerquellen und Grenzen herausgreifen, die – nicht nur von Studierenden, sondern generell von computerlinguistisch nicht vorgebildeten Korpusnutzern – oft als „negative Überraschung“ empfunden werden und die deshalb aus unserer Erfahrung explizit thematisiert werden sollten:

Grenzen der lemmabasierten Abfrage: Studierende begreifen schnell, warum es in einer flektierenden Sprache wie dem Deutschen von Vorteil ist, in einem Korpusrecherchesystem zwischen einer lemmabasierten und einer formbasierten Suche unterscheiden zu können. Sie gehen allerdings davon aus, dass die Trefferliste einer lemmabasierten Abfrage alle Formen des Paradigmas enthält (optimale Ausbeute) und dass keine nicht zum Paradigma gehörigen Formen als Treffer gelistet werden (optimale Präzision). Leider trifft dies nicht immer zu: Probleme mit der Präzision gibt es beispielsweise bei Lemmata, deren Flexionsparadigmen sich überschneiden (z.B. *raten* und *geraten*; *gelingen* und *gelangen*). Wenn man diesen Falltyp kennt, kann man antizipieren, bei welchen Lemmata Probleme auftreten könnten und Strategien entwickeln, um durch spezifischere Abfragen die Präzision zu erhöhen. Probleme mit der Ausbeute gibt es generell bei den deutschen Partikelverben (*einbringen, aufstehen, abtreten* etc.): Die lemmabasierte Suche nach Partikelverben in Korpora erfasst nämlich nur die Formen, in denen Partikel und Verb zusammengeschrieben sind (also *einbrachte, eingebracht, einbrächte, einbringst* etc). Belege, in denen Verbpartikel und

finiter Verbstamm in getrennter Position die Satzklammer bilden (*brachte … ein, bringst … ein*) werden in der Trefferliste nicht angezeigt. Auch hier gilt: Wenn man diese Beschränkung kennt, kann man sie durch geschickte Abfragetechnik kompensieren.

Ob und welche Kompromisse man in Bezug auf die Präzision und Ausbeute einer Abfrage eingehen möchte, hängt letztlich von der verfolgten Fragestellung ab. Wichtig ist es zunächst, sich überhaupt der Möglichkeit bewusst zu sein, dass in den automatisch erzeugten Trefferlisten Pseudotreffer enthalten sind bzw. relevante Belege nicht gefunden werden. Wer z.B. Vorkommensfrequenzen von Verben vergleicht, muss bei der Interpretation seiner Daten berücksichtigen, dass bei einer lemmabasierten Abfrage zu *geraten* auch die Formen von *raten* mitgezählt werden. Umgekehrt gilt wiederum, dass bei einer lemmabasierten Abfrage zu *bringen* auch eine nicht unerhebliche Anzahl von Treffern für Partikelverben wie *einbringen* oder *abbringen* ausgegeben werden, und zwar genau diejenigen, in denen finiter Verbstamm und Partikel in getrennter Position die Satzklammer bilden. Die automatisch generierten Frequenzen können also erheblich von den um Pseudotreffer bereinigten Frequenzdaten abweichen.

Grenzen der Wortartenannotation: Der Nutzwert der automatischen Wortartenannotation kann gerade am Beispiel der deutschen Partikelverben sehr gut verdeutlicht werden: Wer nach Belegen sucht, in denen Partikel und finiter Verbstamm von *einfinden* in getrennter Position im Satz auftreten, kann die Präzision beträchtlich erhöhen, indem er die Suche auf Verwendungen einschränkt, in denen *ein* als trennbare Verbpartikel annotiert ist. Es ist im Allgemeinen kein Problem, die Studierenden mit den Kategorien der Wortartenannotation, z.B. dem Stuttgart-Tübingen-Tagset STTS, vertraut zu machen.[2] Viele Studierende sind jedoch enttäuscht, wenn sie feststellen, dass die automatische Wortartenzuordnung nicht fehlerfrei funktioniert und die prinzipiell sinnvoll formulierten Abfragen dennoch weiterhin viele Pseudotreffer (*false positives*) enthalten. Fehleranfällig sind nicht nur die trennbaren Verbpartikeln, sondern auch viele andere homographe Wortformen, die zu unterschiedlichen Wortarten gehören (z.B. *schicke* als Verb- oder Adjektivform, *sein* als Verb- bzw. Pronomenform, *ja* als Responsiv oder als Abtönungspartikel etc.). Um Frustration zu vermeiden, müssen die Probleme der automatischen Verfahren und Ansätze zur künftigen Optimierung transparent gemacht werden.

„Semantische Blindheit" der Korpusabfragen: Bei vielen linguistischen Analysefragen sucht man eigentlich nach sprachlichen Einheiten in einer bestimmten Bedeutung, also etwa nach *Backfisch* als Bezeichnung für eine Jugendliche oder nach *Ampel* als Bezeichnung für eine Parteienkoalition. Bislang ist eine solche semantische Suche aber in Referenzkorpora nicht möglich, denn sie würde eine automatische Disambiguierung aller mehrdeutigen Wortformen voraussetzen. Davon abgesehen, dass es gar nicht einfach ist, sich auf Kriterien für die Unterscheidung semantischer Lesarten zu einigen – ein Vergleich verschiedener Wörterbücher macht dies deutlich –, gehört die automatische Lesartendisambiguierung (engl. *word sense disambiguation*, *WSD*) leider trotz intensiver Forschungsbemühungen zu den noch nicht befriedigend gelösten Aufgaben in der Computerlinguistik (vgl. den Überblick in RAYSON & STEVENSON 2008). Man kann deshalb derzeit und ggf. auch noch in absehbarerer

[2] Zum Standard und zu Nutzungsbeispielen vgl. STORRER (2011: 223ff.).

Zukunft in großen Korpora nicht automatisch nach bestimmten Wortbedeutungen suchen. Über diese „semantische Blindheit“ der aktuellen Korpustechnologie sind gerade Einsteiger in die korpusgestützte Sprachanalyse enttäuscht, denn sie führt bei der Analyse von stark polysemen Wortformen zu einem hohen manuellen Nachbearbeitungsaufwand. Wer häufiger mit Korpora arbeitet, entwickelt jedoch meist bald ein Gespür dafür, welche Fragen mit welchem Aufwand und mit welchen Suchstrategien bearbeitbar sind. Methodisch muss das Problem der semantischen Blindheit gerade bei der Bewertung statistischer Ergebnisse im Auge behalten werden, denn auch die Statistiken operieren über Formeinheiten und nicht etwa über Bedeutungseinheiten. Wer der Frage nachgeht, ob der Ausdruck *Backfisch* vom Ausdruck *Teenager* verdrängt wird, kann sich nicht allein auf automatisch erzeugte Frequenzverlaufsdiagramme verlassen, sondern muss zunächst die Belege, in denen „Backfisch“ als Bezeichnung für ein Nahrungsmittel verwendet wird, aussondern.

Insgesamt ist es aus unserer Erfahrung wichtig, den Lehramtstudierenden, die ja computerlinguistisch nicht ausgebildet (und meist auch nicht interessiert) sind, zumindest die grundlegenden Verfahrensweisen der automatischen linguistischen Aufbereitung von Korpusdaten transparent und damit die Problemfelder verstehbar zu machen. Wenn dies gelingt, ist nach unserer Erfahrung auch die Bereitschaft vorhanden, die aktuellen Grenzen zu akzeptieren oder diese sogar als Herausforderung zu sehen, um Heuristiken für optimierte, fehlerrobuste Korpusabfragen zu entwickeln. Bereits an kleinen Praxisbeispielen wird ja schnell deutlich, dass auch die noch nicht perfekte digitale Korpustechnologie Analyseoptionen bietet, die durch eine manuell-intellektuelle Analyse von „traditionellen“ Textkorpora in Papierform nicht oder nur mit extrem hohem Aufwand denkbar wären.

4 Einsatz digitaler Sprachressourcen in Lehrveranstaltungen zum Thema „Internetbasierte Kommunikation“

4.1 Gegenstand und Seminarszenario

Die wissenschaftliche Analyse des Sprachgebrauchs und der kommunikativen Besonderheiten in der internetbasierten Kommunikation ist ein Forschungsschwerpunkt der Dortmunder germanistischen Linguistik. Die Befunde, Beschreibungsansätze und Positionen aus diesem Forschungsbereich sind für Studierende der Lehrämter Deutsch sowohl unter fachwissenschaftlicher wie auch fachdidaktischer Perspektive relevant:

(1) Netzgestützte Kommunikationsmöglichkeiten (insbesondere E-Mails, Instant Messaging, Foren, Chats und „soziale Netzwerke“) erfreuen sich in allen Altersgruppen immer stärkerer Beliebtheit, wobei die 14-19-Jährigen die Altersgruppe mit der höchsten Online-Affinität darstellen. Bei der Behandlung sprachlicher Varietäten und sprachlicher Entwicklungstendenzen im Deutschunterricht der Sekundarstufen wird man auf Dauer nicht umhinkommen, diesen Kommunikationsbereich, der gerade für jüngere Nutzergruppen einen selbstverständlichen und wichtigen Teil ihrer alltäglichen Kommunikationswirklichkeit darstellt, systematischer zu berücksichtigen.

(2) Im öffentlichen Diskurs über den Sprachgebrauch im Netz wird häufig die Besorgnis um den Verfall schriftsprachlicher Kompetenzen thematisiert. Wir möchten die Studierenden dazu befähigen, sich in ihrer zukünftigen Rolle als DeutschlehrerInnen und Akteure in

staatlichen Bildungsinstitutionen fachkompetent an diesem Diskurs über die Besonderheiten des Sprachgebrauchs im Internet zu beteiligen. Sie sollen in die Lage versetzt werden, die Auffälligkeiten in der internetbasierten Kommunikation in übergreifende Strömungen der Entwicklung von Schriftlichkeit einzuordnen und funktional zu deuten. Sie sollen auch ein Bewusstsein dafür entwickeln, dass stilistische Variation im Sprachgebrauch kein ausschließlich medienabhängiges Phänomen darstellt (etwa im Sinne einer „Netzsprache" gegenüber der Sprachverwendung außerhalb des Netzes), sondern dass die Sprachverwendung auch *innerhalb* der internetbasierten Kommunikation, abhängig von Zweckbereichen sprachlichen Handelns und unterschiedlichen sozialen/institutionellen Kontexten, z.T. erheblich variiert (z.B. zwischen Chats im Freizeitbereich und Chats in Beratungskontexten, vgl. das Beispiel in Abb. 3).

(3) Darüber hinaus kann und sollte die Thematik auch im schulischen Deutschunterricht (in verschiedenen Klassenstufen) angesprochen werden. Anknüpfungspunkte an die Lehrpläne sind durchaus vorhanden: Der Einfluss technischer Vermittlung auf die Gestaltung von Kommunikationsprozessen ist Gegenstand des Kompetenzziels, „die durch die Informations- und Kommunikationstechnologien bewirkten Veränderungen im Denken, Wahrnehmen und Kommunizieren bewusst" zu machen, das im Lehrplanbereich „Reflexion über Sprache" (vgl. RLP SEK II GYGE DEUTSCH NRW: 25) verankert ist. Im Rahmen des generellen Kompetenzziels „Sprachvarietäten untersuchen und angemessen verwenden können" (RLP SEK II GYGE DEUTSCH NRW: 24) sollten heutzutage auch die unter Punkt (2) angesprochenen schriftsprachlichen Wandelprozesse und sprachlichen Besonderheiten thematisiert werden.

Um der Relevanz der Thematik Rechnung zu tragen, bieten wir an der TU Dortmund regelmäßig fachwissenschaftliche und fachdidaktische Seminare zur internetbasierten Kommunikation an, die im Wahlpflichtbereich der Lehramtsstudiengänge verankert sind. Beispiele für fachwissenschaftliche Seminarthemen sind „Mediale Bedingungen des kommunikativen Handelns", „Internetbasierte Kommunikation", „Linguistische Chat-Forschung" und „Kommunikative Stilistik"; in ihnen werden Konzepte, Forschungsansätze und -ergebnisse zur internetbasierten Kommunikation vermittelt. Im regelmäßig angebotenen fachdidaktischen Seminar „Internetbasierte Kommunikation im Deutschunterricht" bewerten Studierende vorhandene Unterrichtskonzepte zum Thema bzw. stellen eigene Ideen zur Diskussion.

In beiden Bereichen kommen dabei auch digitale Sprachressourcen zum Einsatz. Im fachdidaktischen Kontext dienen sie vor allem als Materialgrundlage für entsprechende Unterrichtseinheiten. Aus dem weiter unten beschriebenen Dortmunder Chat-Korpus lassen sich etwa Datenbeispiele auch aus solchen kommunikativen Handlungsbereichen im Netz gewinnen, die den Schülerinnen und Schülern aus ihrer alltäglichen Online-Nutzung eher weniger vertraut sein dürften (z.B. Chats in Kontexten institutioneller Beratung oder im Kontext politischer Information und Meinungsbildung). Die Gegenüberstellung von Chats aus unterschiedlichen Handlungsbereichen (vgl. Abb. 3) dokumentiert die große Bandbreite der Sprachlichkeit in Chats und kann als Anknüpfungspunkt dienen, um den Zusammenhang von Sprachstil und Faktoren des kommunikativen Settings herauszuarbeiten und auf dieser Basis die Kompetenz zur situativ angemessenen Sprachwahl auch für die neuen netzbasierten Kommunikationsformen zu stärken (vgl. STORRER 2007).

(a)

SPOOKY	Irgendwie ist jetzt an mir was vorbeigeschossen
Findalf	Hausdrache, nö, und ja, er ist scheiß langsam!
Arktikus	GFi: *ggg*...hmm..der aiuch...auff jden Fall zu KArneval *s*
desertstorm betritt den Raum.	
ruebennase langweilt sich immer noch....	
GFi	Karneval in Herne? har..
SPOOKY	Hallo ruebennase, wieso langweilst du dich ?
Hausdrache	Hat jemand ne Ahnung, wie ich CarpeDiem per Mail erreiche??
Arktikus	SPOOKY: so froh, daß Du ein Hausgesit bist und kein menschliches Wesen.....sonst wäre das wohl noch insAuge gegangen..:-)
Arktikus	sei froh..solte es heissen
Findalf	spooky, aha und was war das? sah es aus wie text?*g*
ruebennase	spooky, weil keiner mit mir chattet

(b)

BENUTZER	Können Sie mir sagen, ob das Buch Fn 25665 ausgeliehen ist?
AUSKUNFT	Hallo, wenn Sie einen Moment Geduld haben, schaue ich im Regal nach - Moment
BENUTZER	danke
AUSKUNFT	Ist da, ich lege es Ihnen bei der Information im Erdgeschoss zurück, wenn Sie mirbitte Ihren Namen schreiben.
BENUTZER	Benutzer - bis wann muß ich es abgeholt haben?
AUSKUNFT	Bis wann schaffen Sie es, dann mache ich den entsprechenden Hinweis dran?
BENUTZER	heute oder morgen
AUSKUNFT	O.k. dann schreibe ich bis morgen drauf.
BENUTZER	Vielen Dank!
AUSKUNFT	Gern geschehen und schönen Tag noch.

*** **BENUTZER** hat den chat verlassen. ***

Abbildung 3: Gegenüberstellung von Chats aus unterschiedlichen Handlungsbereichen – zum Beispiel (a) „Plauder-Chat“ und (b) chatbasierte Bibliotheksauskunft.

In den fachwissenschaftlichen Seminaren dienen authentische Datenbeispiele und exemplarische korpusgestützte Untersuchungen als Grundlagen für die (selbstständige) linguistische Analyse und Bewertung sprachlicher, interaktionaler und sozialer Besonderheiten bei der Kommunikation mit Online-Medien. Die Studierenden bearbeiten seminarbegleitend kleine Analyseprojekte auf der Grundlage von Sets authentischer Sprachdaten. Die den Analyseprojekten zugrunde liegenden Forschungsfragen und Untersuchungsdesigns werden dabei entweder vorgegeben oder auf der Basis eigener Ideen gemeinsam mit den Studierenden entwickelt und eingegrenzt. Die Projektarbeit wird unterstützt durch eine Einführung in die genutzten Korpora sowie Hilfestellungen und Beratungsmöglichkeiten durch im Umgang mit den genutzten Ressourcen geschulte TutorInnen. Einzelne Analyseprojekte aus den Seminaren wurden in der Folge im Rahmen von Bachelor-, Master- und Staatsarbeiten aus-

gebaut, etwa zum Sprachstil und zur sprachlichen Variation in der deutschsprachigen *Wikipedia* (Vergleich von Artikel- und Diskussionsseiten), zur sprachlichen Variation in der Chat-Kommunikation und ihrer Didaktisierung, zu den Funktionen von Nicknames in der internetbasierten Kommunikation, zu Leserkommentaren im Online-Journalismus und zum Phänomen des „Splittings" im Chat.

4.2 Genutzte Ressourcen / Dortmunder Chat-Korpus

Im Vergleich zu dem in Abschnitt 3 dargestellten Seminarkonzept stehen in den Seminaren zum Bereich „Internetbasierte Kommunikation" die digitalen Sprachressourcen nicht im Vordergrund, sondern werden punktuell genutzt: als Fundgrube zur Illustration bestimmter Phänomene, als Grundlage für kleine und größere Analyseaufgaben, als empirische Basis für hypothesenerkundende und hypothesenprüfende Untersuchungen zu themenbezogenen Aufgabenstellungen. Da die online verfügbaren Korpussammlungen zur deutschen Sprache (die Korpussammlungen am IDS in Mannheim, die DWDS-Korpora) den Bereich der internetbasierten Kommunikation bislang nicht erfassen, müssen dazu Spezialkorpora genutzt (vgl. BEIßWENGER 2007b, BEIßWENGER & STORRER 2008) bzw. eigene Datensammlungen angelegt werden, z.B. zu Leserkommentaren im Online-Journalismus oder zur Beitragsproduktion in Instant-Messaging-Dialogen. Für Analysen zur Interaktionsorganisation in der Chat-Kommunikation wurde in verschiedenen Seminaren weiterhin eine Sammlung von Transkripten zu Nutzeraktivitäten beim Chatten eingesetzt, die auf einem im Rahmen von BEIßWENGER (2007a) erhobenen Set von Screen-Capturing- und Videodaten basiert.[3]

Da ein Schwerpunkt der Dortmunder Forschung zur internetbasierten Kommunikation im Bereich der Chat-Kommunikation liegt (vgl. BEIßWENGER (Hrsg.) 2001, BEIßWENGER & STORRER (Hrsg.) 2005), wurde in den Jahren 2002–2008 am Lehrstuhl ein Korpus mit Chat-Mitschnitten aufgebaut, das 140.000 Nutzerbeiträge bzw. 1,06 Millionen Tokens aus unterschiedlichen sozialen Handlungsbereichen umfasst: Neben Webchats und IRC-Chats im Freizeitbereich dokumentiert das *Dortmunder Chat-Korpus* Chats in Lehr-/Lernkontexten, in verschiedenen Formen institutioneller Beratung und in journalistischen Nutzungskontexten. Das Korpus wurde nicht nur als empirische Basis für die eigene Forschung (STORRER 2007, LUCKHARDT 2009, BEIßWENGER i.Dr.), sondern auch im Rahmen der o.g. Seminare genutzt.

Die für das Korpus erhobenen Mitschnittsdaten wurden zunächst automatisch bereinigt und in ein einheitliches XHTML-Basisformat konvertiert. Anschließend wurden sie in mehreren semiautomatisch und manuell durchgeführten Aufbereitungsschritten in ein XML-Format überführt, das die Struktur der Mitschnitte und der einzelnen Nutzerbeiträge modelliert, die Nutzerbeiträge klassifiziert, systemgenerierte von nutzergenerierten Primärdatensegmenten unterscheidbar macht und ausgewählte Stilelemente internetbasierter Kommunikation auszeichnet. Hierzu gehören z.B. Emotikons und Inflektive sowie Formen, mit denen Chatter ihre Beiträge an andere Chatter adressieren (vgl. zur Adressierung z.B. BEIßWENGER 2000: 79ff.).

[3] Vier Beispiel-Transkripte aus dem Transkriptkorpus stehen online unter http://www.michael-beisswenger.de/sprachhandlungskoordination/ zur Verfügung.

Forschungsfrage: **Variiert der Gebrauch von Interjektionen und Abtönungspartikeln in unterschiedlichen Typen von Chats?**

(1) Orientieren Sie sich in verschiedenen Grammatiken (DUDEN-Grammatik, IDS-Grammatik etc.) über Formen und Funktionen von **Abtönungspartikeln** (wie z.B. *aber*, *denn*, *vielleicht*, *bloß*, *halt*, …) und **Interjektionen** (wie z.B. *hm*, *ah*, *ach*, *oh*, *äh*, …).

(2) Suchen Sie (manuell) alle Abtönungspartikeln und Interjektionen, die in den beiden Beispiel-Mitschnitten 1+2 vorkommen, die Sie aus dem Ordner „Materialien" im Dateibereich der *Stud.IP*-Plattform herunterladen können. *ACHTUNG:* Zu einigen Partikeln und Interjektionen (z.B. *aber*, *denn*, *oder*, …) gibt es homonyme (gleichlautende) Wörter in anderen Wortartenklassen! Erstellen Sie eine Übersicht, welche Interjektionen und Abtönungspartikeln in den beiden Mitschnitten wie oft vorkommen.

(3) Durchsuchen Sie anschließend (automatisch) mit dem Suchwerkzeug *STACCADo* die folgenden Teilkorpora des Dortmunder Chat-Korpus nach Vorkommnissen derjenigen Wortformen, die Sie in den beiden Beispiel-Mitschnitten als Interjektionen und Abtönungspartikeln vorgefunden haben: (a) alle Plauder-Chats in Medienkontexten; (b) alle Beratungschats der UB Dortmund; (c) die Expertenchats im Hochschulkontext; (d) alle Teilkorpora, die moderierte Chats mit Politikern und Prominenten enthalten.

Berücksichtigen Sie bei der automatischen Suche auch mögliche Schreibvarianten (z.B. *hmmm*, *oda*, …).

Lassen Sie sich von *STACCADo* alle Fundstellen in den durchsuchten Teilkorpora inklusive eines Kontextausschnitts von 10 Beiträgen vor und nach der Fundstelle ausgeben, speichern Sie das Ergebnis in einer Datei und eliminieren Sie manuell Pseudotreffer (z.B. Konjunktion *denn*, Adverb *vielleicht*, …).

(4) Ermitteln Sie für jede Interjektion / Abtönungspartikel, wie häufig sie in den durchsuchten Teilkorpora insgesamt auftritt (einmal in absoluten Zahlen, einmal in Form einer Angabe zur Frequenz pro 100 nutzergenerierten Tokens).

(5) Falls die Verteilung in den durchsuchten Teilkorpora unterschiedlich ist, überlegen Sie Gründe, woran das liegen könnte.

Abbildung 4: Analyseaufgabe einer Projektgruppe im Hauptseminar „Linguistische Chat-Forschung" (SS 2007).

Ein 383 Dokumente (~550.000 Tokens) umfassender, teilweise anonymisierter Ausschnitt des Korpus („Releasekorpus") wird seit 2005 zusammen mit einem spezialisierten, Java-basierten Suchwerkzeug (*STACCADo*) unter *http://www.chatkorpus.tu-dortmund.de* zur freien wissenschaftlichen Nutzung bereitgestellt. Die in das Korpus eingebrachten Annotationen erlauben das Ausblenden von Systemmeldungen bei der Korpusrecherche, die gezielte Suche nach bestimmten (oder innerhalb bestimmter) Typen von Nutzerbeiträgen (Beiträge im Standard-Modus vs. „action messages") oder auch innerhalb der Beiträge einzelner Nutzer. Darüber hinaus können die im Korpus annotierten Stilelemente automatisch ausgefiltert, gezählt oder mit variabel zugeschnittenen Kotexten ausgegeben werden. Für jeden in einem Mitschnitt als Beitragsproduzent dokumentierten Nutzer sind darüber hinaus verschiedene Typen von Metadaten erfasst, u.a. das vermutete Geschlecht sowie die Anzahl der von ihm produzierten Beiträge und Tokens. Auf Basis dieser Metadaten lassen sich mit *STACCADo* automatisch statistische Übersichten zu einzelnen Teilkorpora oder Korpusdo-

kumenten erzeugen: Sog. „Logfile-Profile“ machen verschiedene Teilkorpora hinsichtlich der durchschnittlichen Länge von Chat-Beiträgen oder der Frequenz der Verwendung ausgewählter Sonderelemente wie z.B. Emotikons vergleichbar; „Chatter-Profile“ liefern Übersichten zu den Anteilen einzelner Chatter am Beitragsaufkommen in den jeweils ausgewählten Teilkorpora sowie zum Verhältnis ihrer durchschnittlichen Beitragslängen zur Beitragslänge eines automatisch errechneten „Durchschnitts-Chatters“ und geben für jeden im ausgewerteten Korpusteil bezeugten Chatter das vermutete Geschlecht aus.
Neben der XML-Version, die zur Nutzung zusammen mit dem Suchwerkzeug *STACCADo* heruntergeladen werden muss, ist der online angebotene Korpusausschnitt auch in einer HTML-Version verfügbar, die die 383 Dokumente direkt im Browser darstellbar macht. Auf diese Weise werden die Daten auch für Lehrkräfte und andere am Thema Interessierte zugänglich, die das Korpus nur als Fundgrube für Unterrichtsmaterial nutzen möchten und keine systematischen Auswertungen anstreben. Das Chat-Korpus wurde in 2009 in das Kerncurriculum Deutsch für die gymnasiale Oberstufe des Landes Niedersachsen als Ressource für den Unterricht im Wahlpflichtmodul 2 „Die deutsche Sprache unter dem Einfluss der Neuen Medien“ aufgenommen (vgl. KC DEUTSCH GYGE NI: 51). Ein Beispiel für eine Aufgabenbeschreibung zu einem Analyseprojekt auf der Grundlage des Korpus aus dem Seminar „Linguistische Chat-Forschung“ ist in Abb. 4 wiedergegeben.

4.3 Erfahrungen und Herausforderungen

Nach unseren Erfahrungen mit daten- bzw. korpusgestützten Analyseprojekten in Lehrveranstaltungen zur internetbasierten Kommunikation stehen Studierende einer selbstständigen Bearbeitung kleiner Forschungsfragen auf der Grundlage authentischer Sprachdaten durchaus aufgeschlossen gegenüber. Die entsprechenden Projekte werden meist durchaus engagiert bearbeitet – vorausgesetzt, der praktische Umgang mit der Nutzung des entsprechenden Korpus oder Datensets wurde zuvor eingeführt und es wurden Hinweise (und Begründungen) zur Bewertung der bei der automatischen Korpusrecherche ermittelten Trefferlisten gegeben.

Als sehr sinnvoll hat sich das Angebot einer Tutoren-Sprechstunde speziell zu technischen Fragen (zum Umgang mit der Korpusschnittstelle, zur Abfragesyntax etc.) erwiesen. Hierfür bedarf es TutorInnen, die einerseits über die erforderlichen technischen Kompetenzen verfügen, andererseits selbst bereits grundlegende Erfahrungen mit der Durchführung korpusgestützter Analyseprojekte gesammelt haben. Mit einer Verstetigung entsprechender Veranstaltungen im Fach können Studierende mit dem erforderlichen Kompetenzprofil aus den Seminaren selbst rekrutiert und in Folgeseminaren als TutorInnen eingesetzt werden. Auch die Möglichkeit, vor der eigentlichen Präsentation der Projektergebnisse im Seminarplenum (z.B. in der Semestermitte) Zwischenergebnisse zu diskutieren und Fragen, die sich bei der praktischen Durchführung der Analysen ergeben haben, zu besprechen, hat sich bewährt, lässt sich aber i.d.R. nur in kleineren Seminaren sinnvoll realisieren.

Bei den Korpora, die derzeit zum Bereich der internetbasierten Kommunikation existieren, handelt es sich entweder um reine Rohdatenkorpora, deren Inhalte zu Zwecken einer einheitlichen Verwaltung bestenfalls formal vereinheitlicht wurden, ansonsten aber keiner weiteren Aufbereitung speziell in Hinblick auf linguistische Analysezwecke unterzogen

wurden (vgl. BEIßWENGER 2007b, BEIßWENGER & STORRER 2008). Die wenigen Korpora, deren Datenbestand entsprechend aufbereitet wurde (z.B. das Dortmunder Chat-Korpus), sind – zumindest zu großen Teilen – *hand*annotiert. Im Vergleich zu den in Abschnitt 3 erwähnten linguistisch aufbereiteten Textkorpora gibt es keine Lemmatisierung und auch kein Wortartentagging. Dies hat damit zu tun, dass Werkzeuge für die automatische linguistische Aufbereitung von Sprachdaten (Tokenisierer, Lemmatisierer, Part-of-speech-Tagger, Chunk Parser) meist an redigierter Schriftsprache (oft Zeitungskorpora) trainiert sind und sich bei den wenig normgerechten Chat-Daten nur begrenzt bewähren. Eine Anpassung dieser Werkzeuge an Abweichungen von der (gerade in Zeitungskorpora i.d.R. sehr mustergültig eingehaltenen) orthographischen Norm, die nicht nur für Chats, sondern für viele informelle Nutzungskontexte der internetbasierten Kommunikation typisch sind, muss erst noch geleistet werden – diese Anpassung ist ein wichtiges Desiderat in Bezug auf die Verarbeitung von Sprachdaten aus dem Netz und ihre Repräsentation in Korpora. Entsprechend steht die Möglichkeit einer Disambiguierung formgleicher Ausdrücke durch Einbeziehung etwa von Wortartenkategorien oder eines Grundformenoperators in die Suchanfragen für solche Korpora nicht zur Verfügung. Je nachdem, welche Kategorien im Korpus (semiautomatisch oder manuell) annotiert wurden und welche nicht, muss also sowohl für die Ermittlung der für eine Analysefrage relevanten Treffer wie auch für die intellektuelle Nachsortierung der für die Anfrage faktisch erhaltenen Treffer ein ungleich größerer Aufwand betrieben werden als im Falle der Arbeit mit linguistisch aufbereiteten Textkorpora. Deshalb ist es gerade für die Arbeit mit Spezialkorpora zur internetbasierten Kommunikation wichtig, ein Verständnis dafür zu vermitteln, welche Arten von Fragestellungen sich auf der Grundlage der im Korpus enthaltenen Daten und Annotationen direkt beantworten lassen und für welche Arten von Fragen die ermittelten Treffer nur „rohe" Ergebnisse darstellen, in deren Bewertung noch weitere intellektuelle Nacharbeit investiert werden muss.

5 Desiderate und Perspektiven

Die Behandlung von und der systematische Umgang mit digitalen Sprachressourcen ist derzeit zwar noch kaum explizit in den Lehrplänen für Schulen und den fächerspezifischen Bestimmungen für die Lehramtsstudiengänge mit Deutsch verankert, lässt sich aber – wie in Abschnitt 2 gezeigt wurde – sehr gut an verschiedene der dort formulierten Kompetenzziele anschließen. Für die Nutzung digitaler lexikalischer Ressourcen zur deutschen Sprache (Wörterbücher, Wörterbuchportale und digitale lexikalische Informationssysteme) werden allerdings erweiterte Kompetenzen zur Bewertung von Verlässlichkeit und Qualität benötigt, die in der „prädigitalen" Wörterbuchdidaktik noch kaum thematisiert sind (vgl. BEJOINT 1989). Wie wir in den Abschnitten 3 und 4 an Beispielen gezeigt haben, bieten digitale Textkorpora zwar viele interessante Anknüpfungspunkte für die fachwissenschaftliche und fachdidaktische Ausbildung in den Lehramtsstudiengängen; ob und wieweit die aktuell verfügbaren Online-Korpora zur deutschen Sprache schon direkt im Schulunterricht eingesetzt werden können, muss jedoch noch weiter erprobt werden.

Aus unseren Erfahrungen heraus könnten die folgenden Faktoren dazu beitragen, den Umgang mit digitalen Korpusressourcen auch für computerlinguistisch nicht vorgebildete Nutzer – z.B. im Lehramtsstudium und in der Schule – noch attraktiver zu machen:

Einheitliche Schnittstellen und Anfragesprachen: Bislang muss für fast jede Korpusumgebung eine eigene Abfragesyntax und der Umgang mit einer spezifisch auf dieses Korpus zugeschnittenen Nutzerschnittstelle erlernt werden. Dies erschwert es, Daten aus unterschiedlichen Korpusbeständen zu nutzen oder gar systematisch zu vergleichen (z.B. Text- vs. Gesprächskorpora, Spezialkorpora vs. Referenzkorpora). Ein Desiderat wäre also eine einheitliche Oberfläche für mehrere Korpora sowie die Vereinheitlichung von Abfragesprachen.

Verständliche Dokumentation von Annotationen und von automatisch erzeugten Angaben: Die in ein Korpus eingebrachten Annotationen (Tagsets) und zugrunde liegenden Analysekategorien sollten so dokumentiert sein, dass sie auch für Nutzer ohne computerlinguistische Vorbildung verständlich und nachvollziehbar sind. Wenn in den Projekten die Zeit für eine solche Dokumentation fehlt, sollte man zumindest eine Web2.0-Funktion anbieten, in denen sich die Nutzer wechselseitig austauschen und unterstützen können. Auch die Funktionen, die automatische Angaben aus den Korpusdaten erzeugen (Frequenzangaben, Kookkurrenz- bzw. Wortprofile, Kookkurrenzgraphen), sollten ausführlich dokumentiert werden, um Fehlinterpretationen der Daten zu vermeiden.

Grenzen und Problemfelder offen ansprechen: Grenzen der aktuellen Korpustechnologie sollten in den Hilfetexten und Tutorials zum Umgang mit Korpora offen angesprochen und erklärt werden (z.B. Probleme und typische Fehler bei der automatischen Lemmatisierung und Wortartenannotation oder das Problem der „semantischen Blindheit“ der Korpussuche). Eine Kenntnis der Grenzen automatischer Sprachverarbeitung hilft Nutzerinnen und Nutzern ohne computerlinguistische Grundkenntnisse, die Potenziale von Korpora realistisch einschätzen und für die korpusgestützte Bearbeitung linguistischer Forschungsfragen geeignete Recherchestrategien entwickeln zu können. Wie in Abschnitt 3 erläutert, haben wir die Erfahrung gemacht, dass Studierende diesen Grenzen tolerant gegenüberstehen und sogar Spaß daran haben, nach Abfragen mit möglichst guter Präzision und Ausbeute zu suchen. Voraussetzung ist, dass sie verstehen, wie die Annotationen zustande kommen und warum die automatisch erzeugten Annotationen oft nicht fehlerfrei sind. Wenn Studierende hingegen ohne Einführung eigenständig Korpora nutzen, sind sie über auftretende Probleme schnell enttäuscht und bewerten die Ressource vorschnell als irrelevant bzw. unbrauchbar.

Hinweise zu Strategien bei der Korpusabfrage: Sehr hilfreich für Nutzerinnen und Nutzer mit keinen oder nur geringen korpuslinguistischen Grundkenntnissen sind Tipps für die Reduktion von Pseudotreffern und für den Umgang mit typischen Problemfällen bei der Korpusrecherche (z.B. orthographische Varianten, Homographen, Partikelverben). Auch diese könnten ggf. über Web2.0-Angebote auch von den Nutzern selbst erstellt werden.

Arbeitsplatz für die Weiterarbeit mit Rechercheergebnissen: Aus den vom Korpusrecherchesystem erzeugten Trefferlisten ergibt sich in aller Regel noch nicht die Antwort auf eine linguistische Forschungsfrage. Um die Trefferliste in eine Liste mit „echten“ Belegen für das gesuchte Phänomen zu überführen, ist es in den allermeisten Fällen erforderlich, die Treffer weiterzubearbeiten, z.B. indem Pseudotreffer eliminiert werden. In weiteren Bearbeitungsschritten können Subklassen gebildet oder weitere Annotationen hinzugefügt werden. Bislang erfolgen diese Schritte der Datenbearbeitung meist unabhängig von der Korpusumgebung (d.h. nach dem Export der Daten in eine Datei) mit einem externen Werkzeug (einer Annotationsumgebung, mit Excel o.ä.). Dadurch lassen sich die weiterbearbeiteten

Daten allerdings auch nachträglich i.d.R. nicht mehr ohne Weiteres mit den Korpusdokumenten vernetzen, aus denen sie extrahiert wurden. Häufig ergibt sich aber gerade bei der Weiterarbeit mit einmal aus einem Korpus gewonnenen Treffern die Notwendigkeit, diese erneut mit dem Korpus in Beziehung zu setzen – beispielsweise, um zu einzelnen „problematischen" Treffern oder Belegen größere Kotexte einzusehen als diejenigen, die beim Export der betreffenden Treffer gewählt oder automatisch mitausgegeben wurden. Ein weiteres Desiderat für die Zukunft wäre deshalb ein „virtueller" korpuslinguistischer Arbeitsplatz, der die Nutzer nicht nur bei der Abfrage des Korpus und beim Export ihrer Trefferlisten unterstützt, sondern ihnen auch die Weiterarbeit mit den Daten direkt in der Korpusumgebung ermöglicht und/oder Funktionen anbietet, mit denen sich Treffer auch nach ihrer Weiterbearbeitung in externen Werkzeugen wieder im Korpus ermitteln lassen. Ein solcher korpuslinguistischer Arbeitsplatz sollte idealiter auch häufig benötigte Funktionen zur quantitativen Auswertung von Belegsammlungen anbieten. Die Anreicherung von Korpus-Benutzerschnittstellen um solche und weitere Funktionen würde diese schrittweise von Abfrage- und Exportschnittstellen zu *Umgebungen für korpuslinguis-tisches Arbeiten* weiterentwickeln, in denen das Korpus nicht lediglich als *Ressource* präsentiert, sondern – orientiert an typischen linguistischen Nutzungsszenarien – als *Hilfsmittel* des wissenschaftlichen Prozesses dargeboten wird.

Für die Umsetzung dieser Desiderate wäre es sicherlich hilfreich, wenn die Entwickler von Korpustechnologie, die bislang ja vornehmlich aus dem Bereich der Computerlinguistik oder der linguistisch orientierten Informatik kommen, den Kontakt mit verschiedenen Gruppen von Korpusnutzern intensivieren. Angehende Lehrerinnen und Lehrer sind neben anderen Nutzergruppen (etwa Journalisten, Übersetzern und Sprachmittlern) sicherlich interessante Multiplikatoren für die vielfältigen Möglichkeiten, die digitale Sprachressourcen und Sprachkorpora bieten.

6 Literatur und erwähnte Sprachressourcen

6.1 Literatur

ALBERT, RUTH & COR J. KOSTER (2002). Empirie in Linguistik und Sprachlehrforschung. Ein methodologisches Arbeitsbuch. Tübingen.

BEIßWENGER, MICHAEL (2000). Kommunikation in virtuellen Welten: Sprache, Text und Wirklichkeit. Stuttgart.

BEIßWENGER, MICHAEL (2007a). Sprachhandlungskoordination in der Chat-Kommunikation. Berlin. New York (Linguistik – Impulse & Tendenzen 26).

BEIßWENGER, MICHAEL (2007b). Corpora zur computervermittelten (internetbasierten) Kommunikation. In: Zeitschrift für germanistische Linguistik 35, 496-503.

BEIßWENGER, MICHAEL (im Druck). Raumorientierung in der Netzkommunikation. Korpusgestützte Untersuchungen zur lokalen Deixis in Chats. In: Barbara Frank-Job, Alexander Mehler & Tilmann Sutter (Hrsg.): Die Dynamik sozialer und sprachlicher Netzwerke. Wiesbaden.

BEIßWENGER, MICHAEL (Hrsg., 2001). Chat-Kommunikation. Sprache, Interaktion, Sozialität & Identität in synchroner computervermittelter Kommunikation. Perspektiven auf ein interdisziplinäres Forschungsfeld. Stuttgart.

BEIßWENGER, MICHAEL & ANGELIKA STORRER (2008). Corpora of Computer-Mediated Communication. In: Lüdeling & Kytö (eds.), 292-308.

BEIßWENGER, MICHAEL & ANGELIKA STORRER (Hrsg., 2005). Chat-Kommunikation in Beruf, Bildung und Medien: Konzepte – Werkzeuge – Anwendungsfelder. Stuttgart.

BEJOINT, HENRY (1989). The Teaching of the Dictionary Use: Present State and Future Tasks. In: Franz Josef Hausmann, Oskar Reichmann, et al. (Hgg.): Wörterbücher. Ein internationales Handbuch zur Lexikographie, 1. Teilband. Berlin. New York (HSK 5.1), 208-215.

BORTZ, JÜRGEN & NICOLA DÖRING (2006). Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler. 4., überarb. Aufl. Berlin u.a.

[FSB GERM BFP DORTMUND] Fächerspezifische Bestimmung für das Fach Germanistik zur Prüfungsordnung für den Bachelor-Studiengang mit fachwissenschaftlichem Profil im Modellversuch „Gestufte Studiengänge in der Lehrerbildung". Amtliche Mitteilungen der Universität Dortmund 12/2007, 34-43. URL: http://www.tu-dortmund.de/uni/studierende/pruefungsangelegenheiten/ord/FSB/FSB_Bachelor/15_FSB_Germanistik_BfP.pdf.

ENGELBERG, STEFAN & LOTHAR LEMNITZER (2009). Lexikographie und Wörterbuchbenutzung. 4., überarbeitete Auflage. Stauffenburg: Tübingen.

EISENBERG, PETER & WOLFGANG MENZEL (1999). Grammatik-Werkstatt. In: Praxis Deutsch 22, Heft 129, 14-26.

GEYKEN, ALEXANDER (2007). The DWDS corpus: A reference corpus for the German language of the 20th century. In: Christiane Fellbaum (ed.): Collocations and Idioms, London, 23-40.

HOFFMANN, LUDGER (2006). Funktionaler Grammatikunterricht. In: Tabea Becker & Corinna Peschel (Hrsg.): Gesteuerter und ungesteuerter Grammatikerwerb. Baltmannsweiler, 20-45.

[KC DEUTSCH GYGE NI] Kerncurriculum Deutsch für das Gymnasium/gymnasiale Oberstufe, die Gesamtschule/gymnasiale Oberstufe, das Fachgymnasium, das Abendgymnasium, das Kolleg. Hrsg. v. Niedersächsischen Kultusministerium 2009. WWW-Ressource: http://db2.nibis.de/1db/cuvo/datei/kc_deutsch_go_i_2009.pdf

LEMNITZER, LOTHAR & HEIKE ZINSMEISTER (2006). Korpuslinguistik: Eine Einführung. Tübingen.

LUCKHARDT, KRISTIN (2009). Stilanalysen zur Chat-Kommunikation. Eine korpusgestützte Untersuchung am Beispiel eines medialen Chats. Diss., TU Dortmund. Digitale Ressource: http://hdl.handle.net/2003/26055.

LÜDELING, ANKE & MERJA Kytö (Eds., 2008/2009). Corpus Linguistics. An International Handbook. 2 Bde. Berlin. New York (HSK 29.1/29.2).

MCENERY, TONY, RICHARD XIAO & YUKIO Tono (2006). Corpus-Based Language Studies – an advanced resource book. London. New York.

RAYSON, PAUL & MARK STEVENSON (2008). Sense and semantic tagging. In: Lüdeling & Kytö (eds.), 564-578.

[RLP SEK II GYGE DEUTSCH NRW] Richtlinien und Lehrpläne für die Sekundarstufe II – Gymnasium/Gesamtschule in Nordrhein-Westfalen. Deutsch. Hrsg. v. Ministerium für Schule und Weiterbildung, Wissenschaft und Forschung 1999.

STORRER, ANGELIKA (2007). Chat-Kommunikation in Beruf und Weiterbildung. In: Der Deutschunterricht 1/2007, 49-61.

STORRER, ANGELIKA (2010). Deutsche Internet-Wörterbücher: Ein Überblick. In: Lexicographica. Internationales Jahrbuch für Lexikographie. Vol. 27, 155–164.

STORRER, ANGELIKA (2011). Korpusgestützte Sprachanalyse in Lexikographie und Phraseologie. In: Karlfried Knapp u.a. (Hrsg.): Angewandte Linguistik. Ein Lehrbuch. 3. Auflage. Tübingen, 216-239. [Aufgaben und Lösungen zum Artikel online unter http://www.studiger.tu-dortmund.de/images/Korpuslinguistik_aufgaben.pdf].

6.2 Erwähnte Ressourcen

[CANOO.NET]: Portal zur deutschen Sprache (Wörterbücher und Grammatik): http://www.canoo.net/

[DICT.CC] Verbund multilingualer Wörterbücher mit Deutsch als Äquivalentsprache: http://browse.dict.cc/

DORTMUNDER CHAT-KORPUS: http://www.chatkorpus.tu-dortmund.de

[DWDS]: Verbund digitaler Wörterbücher und Textkorpora (Berlin-Brandenburgische Akademie der Wissenschaften): http://www.dwds.de

[DWDS-WDG]: Digitales Wörterbuch der deutschen Sprache auf der Basis des digitalisierten „Wörterbuch der deutschen Gegenwartssprache“ (Berlin-Brandenburgische Akademie der Wissenschaften): http://www.dwds.de

[DWB-ONLINE]: Das Deutsche Wörterbuch von Jacob und Wilhelm Grimm Online: http://www.dwb.uni-trier.de

[ELEXIKO]: elexiko – Online-Wörterbuch zur deutschen Gegenwartssprache: http://www.ids-mannheim.de/lexik/elexiko/

[GRAMMIS]: Grammis – das grammatische Informationssystem des Instituts für deutsche Sprache: http://hypermedia.ids-mannheim.de/grammis/

[OPENTHESAURUS]: Wörterbuch für Synonyme und Assoziationen (deutsch): http://www.openthesaurus.de/

[OWID]: Verbund digitaler Wörterbücher (Institut für deutsche Sprache): http://www.owid.de/

[PROGR@MM]: ProGr@mm – die propädeutische Grammatik: http://hypermedia.ids-mannheim.de/programm/

[SZENESPRACHENWIKI]: Duden: Neues Wörterbuch der Szenesprachen: http://szenesprachenwiki.de/

[WIKTIONARY]: Wiktionary: das freie Wörterbuch (deutsch): http://de.wiktionary.org/wiki/Wiktionary:Hauptseite

[WORTWARTE]: Die Wortwarte: http://www.wortwarte.de/

Noah Bubenhofer

Korpuslinguistik in der linguistischen Lehre: Erfolge und Misserfolge

Für die sprachwissenschaftliche Ausbildung an den Universitäten ist es zwar unabdingbar, die Studierenden in die Theorie und Methoden der Korpuslinguistik einzuführen, doch als Lehrperson kämpft man dabei mit einer Reihe von Problemen, denn das technische und methodische Know-how der Studierenden ist oft sehr heterogen. Zudem zeigt sich die Wichtigkeit, die Studierenden für korpuslinguistisches Arbeiten begeistern zu können, indem sie an attraktives Anschauungsmaterial herangeführt werden. Im Folgenden zeige ich an einigen Beispielen, welche Themen in den Bereichen Semantik, Textlinguistik, Diskurs- und der Kulturanalyse sinnvollerweise korpuslinguistisch bearbeitet werden können. Zudem versuche ich anhand des Nutzungsverhaltens meiner Online-Einführung in die Korpuslinguistik die Bedürfnisse von Anwendern an Methoden und Werkzeuge der Korpuslinguistik abzuleiten.

1 Einleitung

Korpuslinguistische Methoden gehören inzwischen zum Standardrepertoire vieler Forscherinnen und Forscher in der Sprachwissenschaft. So verwundert es nicht, dass die Korpuslinguistik auch Eingang in die linguistische Lehre findet – auch wenn dies erst zaghaft und nicht in allen Teildisziplinen gleichermaßen geschieht.

Dabei müssen allerdings eine Reihe von Hürden überwunden werden:

- Korpuslinguistisches Arbeiten setzt Wissen über empirische Forschung generell voraus. Diese Grundlage muss teilweise erst aufgebaut werden.
- Teilweise fehlt bei den geisteswissenschaftlich geprägten Studierenden das technische Know-how, um korpuslinguistisch arbeiten zu können. Zudem ist das manchmal gepaart mit dem fehlenden Selbstvertrauen, sich dieses Know-how anzueignen.
- Oft bedingen korpuslinguistische Arbeiten einen großen Aufwand, sowohl für Lernende als auch die Betreuenden, der im Rahmen eines Studiums nicht geleistet werden kann.

Trotzdem ist es gewinnbringend, mit Studierenden korpuslinguistisch zu arbeiten. Im Folgenden möchte ich deshalb an einigen Beispielen aus meiner eigenen Lehrpraxis

zeigen, welche Themen in den Bereichen Semantik, Textlinguistik, Diskurslinguistik und Kulturanalyse bearbeitet werden können.

Bei korpuslinguistischem Arbeiten geht es immer auch um technische Aspekte: Welche Ressourcen und Werkzeuge gibt es, die von Anwenderinnen und Anwendern ohne Informatik-Hintergrund leicht eingesetzt werden können? Die Präsentation der Beispiele studentischer Arbeiten wird deshalb ergänzt durch Hinweise auf Ressourcen und Werkzeuge, die sich in der Lehre bewährt haben. Trotzdem bleiben Wünsche an die Entwicklung von Analysewerkzeug offen, die ich skizzieren möchte.

Um diese Wünsche und Bedürfnisse auf eine breitere Basis zu stellen, habe ich die Zugriffe auf die online verfügbare „Einführung in die Korpuslinguistik“ (Bubenhofer, 2006-2011) analysiert und ein Nutzerprofil erstellt, das ich ebenfalls präsentiere.

2 Korpuslinguistik in der Lehre

2.1 Konzepte und Ziele

Meine Erfahrung zu Korpuslinguistik in der Lehre beruht auf folgenden Lehrveranstaltungen und Betreuungsangeboten:

- Kurse „Einführung in die Korpuslinguistik“ im Grundstudium/Bachelor-Studiengang am Deutschen Seminar der Universität Zürich.
- Seminare an den Universitäten Zürich und Mannheim (Hauptstudium/Master-Studiengang) zu den Themen Diskurs-/Kulturanalyse und Semantik (Lexikographie), in denen korpuslinguistisches Arbeiten die methodische Grundlage darstellte.
- Betreuung von Seminar- und Abschlussarbeiten zu verschiedenen Themen mit korpuslinguistischer Methodik.
- Verschiedene Workshops und Beratung für fortgeschrittene Studierende und Doktorierende zu korpuslinguistischen Themen in den Bereichen Semantik, Grammatik, Textlinguistik, Diskurs- und Kulturanalyse.

Das Ziel sowohl der einführenden Kurse als auch der Seminare war, die Studierenden an empirisches Arbeiten mit Korpora heranzuführen. Es geht also darum, den Weg von einer Hypothese zur Operationalisierung und Analyse aufzuzeigen und sich grundlegende Gedanken über den Stellenwert von Korpora als Datengrundlage für Analysen des Sprachgebrauchs zu machen. Darauf aufbauend lernen die Studierenden verfügbare Ressourcen und Werkzeuge kennen und anhand eigener Fragestellungen anzuwenden. Ein typischer Seminarplan einer solchen Einführung ist in Tabelle 1 dargestellt.[1] Dabei bewegt man sich zwischen den folgenden Polen:

[1] Als begleitende Literatur nutzte ich neben themenspezifischer Literatur die Einführungen Lemnitzer/Zinsmeister (2006) und Scherer (2006).

1. **Grundlagen:** Begriffsklärung, korpuslinguistische Denkweise (Arm-Chair Linguist vs. Corpus Linguist), Anwendungen
2. **Empirisches Arbeiten:** Thesenbildung, Operationalisierung
3. **Korpora Grundlagen:** Repräsentativität, Korpusgröße, Korpustypen, Annotation
4. **Bestehende Korpora nutzen:** DeReKo IDS (o. J.), DWDS (o. J.), Baumbanken
5. **Methoden:** Recherchen, Ergebnisdarstellung, Kollokationen, n-Gramme, statistische Auswertungen
6. **Eigene Korpora aufbauen**

Tabelle 1: Typischer Seminarplan „Einführung in die Korpuslinguistik“

Pol 1: Analyseebenen Belege finden ↔ Muster entdecken ↔ systematische statistische Auswertungen durchführen

Die Studierenden lernen verschieden elaborierte Methoden der Korpusanalyse kennen, wobei der einfachste Zugang über klassische, korpusbasierte Analysen erfolgt: Konkordanzen und Belege analysieren und kategorisieren, Kollokationsanalysen, Vergleiche mit anderen Korpora. Ziel ist es aber, fortgeschrittenere Analysen zu machen, mit denen die Beobachtung von Einzelbelegen systematisiert und abstrahiert werden.

Pol 2: Datengrundlage bestehende Korpora nutzen ↔ eigene Korpora aufbauen

Es sind bereits viele sofort nutzbare Korpora verfügbar, die den Einstieg in Korpusanalysen erleichtern. Für viele Forschungszwecke ist es aber sinnvoll, eigene Korpora erstellen zu können.

Pol 3: Know-how viel technisches Know-how ↔ wenig technisches Know-how

Das vorhandene computertechnische Know-how der Studierenden ist meistens sehr heterogen. Deshalb ist es wichtig, auf Studierende mit eher wenig Know-how Rücksicht zu nehmen und sie aber gleichzeitig zu ermutigen, trotzdem anspruchsvolle Analysemethoden auszuprobieren. Oft können Projekte auch in Gruppen durchgeführt werden, in denen sich die Studierenden mit ihren unterschiedlichen Fähigkeiten ergänzen.

Aus diesen Polen leiten sich bereits die Bedürfnisse nach Ressourcen und Werkzeugen her, die idealerweise für die Lehre verfügbar sind, um alle Aspekte abdecken zu können: Man benötigt Software, um allen Analyseebenen in bestehenden aber auch selbst aufgebauten Korpora nachgehen zu können, wobei diese Software sowohl für „Poweruser“ als auch technisch wenig bewanderte Studierende nutzbar sein sollte.[2]

[2] Dies gilt natürlich nicht nur im Bereich der Lehre, sondern genau so in der Forschung.

Wichtig, um die Studierenden zu korpuslinguistischen Arbeiten zu animieren, sind anschauliche Beispiele solcher Arbeiten aus der Forschung.[3] Einerseits machen intuitiv nutzbare Korpusrecherche-Tools wie Wortschatz Leipzig (o. J.), der Google Books Ngram Viewer (o. J.)[4] oder eine anschauliche Anwendung wie die „Wortwarte“ (Lemnitzer, 2011) bereits Lust darauf, Sprachdaten zu analysieren. Andererseits dienen dazu auch Forschungsarbeiten, durchaus populärwissenschaftlich aufbereitet, wie die Analysen zu den vergangenen US- und Bundestagswahlen (Bubenhofer u. a., 2008, 2009), von denen ich als Forschungsgruppenmitglied aus erster Hand berichten konnte.

2.2 Beispiele studentischer Arbeiten

Im Folgenden berichte ich von einigen korpuslinguistischen Arbeiten von Studierenden, die ich betreute.[5] Es geht weniger darum, die genauen Inhalte zu referieren, sondern die eingesetzten Methoden und Werkzeuge zu erwähnen. Die Arbeiten sind naturgemäß unveröffentlicht.

David Papst nimmt sich in **„*klein* und *winzig*. Eine Korpusuntersuchung zur Synonymie“** (Papst, 2005) eine klassische Fragestellung der Semantik vor. Als Basis für die Untersuchung dieser Quasisynonyme dient ihm das DWDS (o. J.), aus dem er eine Zufallsauswahl von Belegen für die Lemmata extrahiert und manuell kategorisiert. Er verzichtet völlig auf Standardverfahren wie eine Kollokationsanalyse, kann die semantisch-funktionalen Differenzen der beiden Lemmata aber trotzdem gut beschreiben.

Technisch und theoretisch etwas avancierter ist eine Arbeit von Igor Matic: **„Konzeptuelle Metaphern der Wirtschaftskrise in der NZZ am Sonntag“** (Matic, 2009). Er stellt sich ein eigenes Korpus aus 53 Zeitungsartikeln zusammen und kategorisiert darin manuell Metaphern. Um die Zeitungsartikel zu durchsuchen verwendet er „AntConc“[6] (vgl. Abbildung 1), eine Konkordanzsoftware, die aber auch Kollokationen, n-Gramme und bei Korpusvergleichen Schlüsselwortanalysen vornehmen kann.

Paul Rauber benutzt in seiner diskurslinguistisch ausgerichteten Arbeit **„Intellektuelle im Diskurs. Zwischen Hybris und Machtkritik“** (Rauber, 2009) mehrere Korpora. Einerseits verwendet er ein eigenes Korpus von Zeitungsartikeln aus dem Schweizer „Tages-Anzeiger“, die er über eine Schlüsselwortsuche mit den Lemmata *Intellektueller/intellektuell* zusammenstellt und mit der Konkordanzsoftware „AntConc“ verwaltet. Als Referenzkorpora benutzt er die „Frankfurter Rundschau“ über die COSMAS II-Schnittstelle des DeReKo IDS (o. J.), das gesamte DWDS (o. J.) und die Kollokationsprofile von Wortschatz Leipzig (o. J.). In allen Korpora nutzt er die Funktionen zum

[3] Dabei spielt die eigene Forschung naturgemäß eine besonders wichtige Rolle, wie bei mir die Arbeit zu korpuslinguistischen Methoden in der Diskurs- und Kulturanalyse (Bubenhofer, 2009).

[4] Der Google Ngram Viewer erzielte einige Aufmerksamkeit durch die populärwissenschaftliche Lancierung des Forschungszweigs „Culturomics“ (Michel u. a., 2011; Lieberman u. a., 2007).

[5] Einen Teil der Arbeiten betreute ich zusammen mit Angelika Linke.

[6] Die Software wurde von Laurence Anthony entwickelt: `http://www.antlab.sci.waseda.ac.jp/antconc_index.html` (23. März 2011), Anthony (2010).

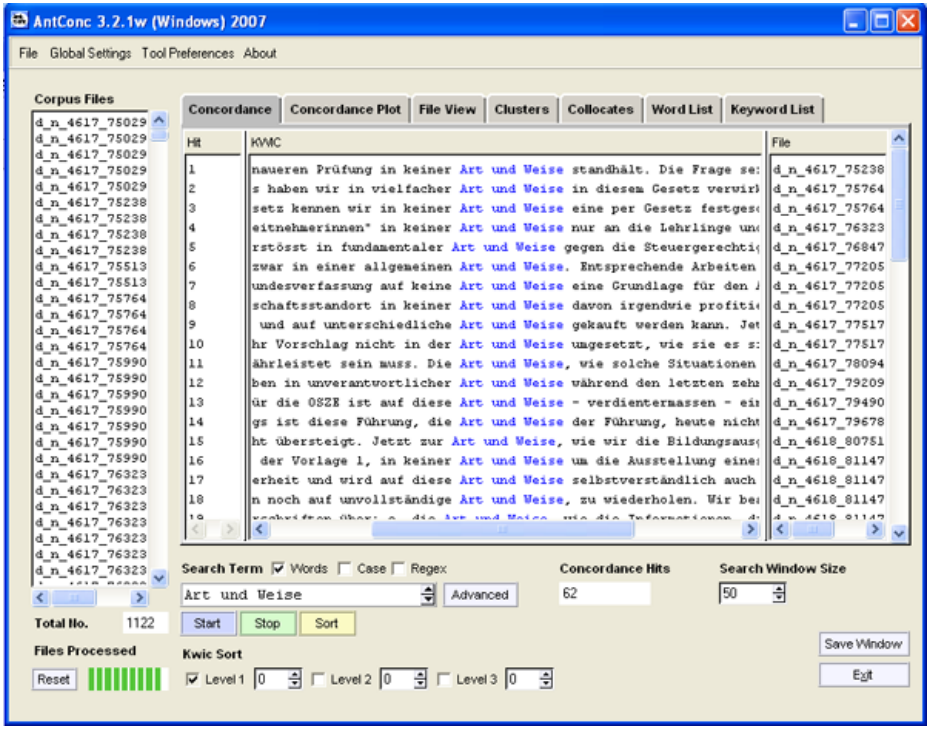


Abbildung 1: „AntConc“ ist eine einfach bedienbare Konkordanzsoftware, mit der auch Kollokationen, n-Gramme und Schlüsselwörter berechnet werden können (Anthony, 2010).

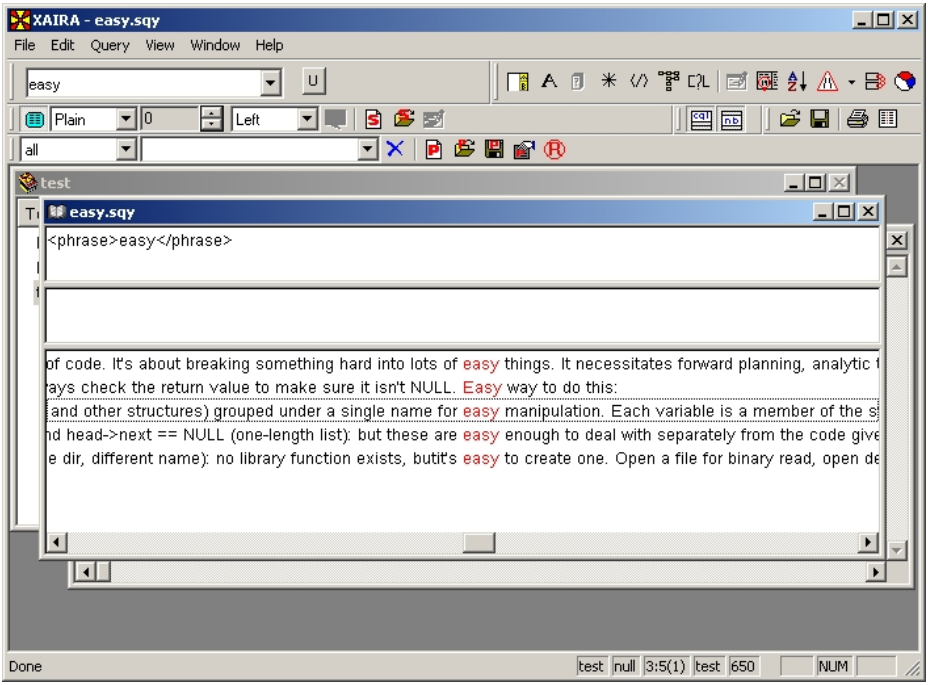


Abbildung 2: „Xaira“ ist eine Konkordanzsoftware, die XML-Dokumente verarbeiten kann (Oxford University Computing Services, 2011).

Berechnen von Kollokationen, um die Verwendungsweisen des Konzeptes *intellektuell* semantisch und diskurslinguistisch zu beschreiben.

In eine ähnliche Richtung zielt Verena Casanas Arbeit **„Homosexualität. Analyse der Paradigmengruppe *homosexuell* – *gleichgeschlechtlich* anhand der taz“** (Casana, 2009). Im selber anhand der Schlüsselwörter zusammengestellten Korpus der „tageszeitung“ von 1994–2008 untersucht sie mit „AntConc“ die Belege und Kollokationen zu den untersuchten Lemmata und deutet den Wandel der Verwendungsweise diskurslinguistisch.

Während die bisher beschriebenen Arbeiten mit relativ wenig Aufwand auskamen, um die Daten für die Analyse vorzubereiten, ist das in Tamara Weibels Arbeit anders: **„*Mieterinnen* oder *Mieter* – *Schweizervolk* oder *Ausländer*? Parteispezifische Personen- und Personengruppenbezeichnungen der SP und SVP im Schweizer Parlament“** (Weibel, 2009). Grundlage sind die Protokolle des Schweizer Parlaments, die allerdings von einem computerlinguistisch bewanderten Kommilitonen aufbereitet werden mussten, um die Redebeiträge zu extrahieren und den jeweiligen

Sprechern und Parteien zuordnen zu können. So aufbereitet nutzt die Autorin ebenfalls „AntConc“, um die darin vorhandene Funktion „Keywords“ zu verwenden, die beim Vergleich zweier Teilkorpora die jeweils statistisch signifikanten Wörter berechnet.[7] Die so als parteitypisch eruierten Lemmata werden dann durch Kollokationsanalysen und eine manuelle Kategorisierung semantisch-funktional gedeutet.

Auch Sara Baertschi bewegt sich mit ihrer Abschlussarbeit **„Der Berg ruft. Sprachgebrauchsmuster von 1920-1945 in der Literatur des Schweizer Alpen-Clubs“** (Baertschi, 2010) im Feld der Diskursanalyse. Basis ist ein Korpus alpinistischer Texte (Text+Berg-Korpus, 2011). Auch sie benutzt „AntConc“, um Belege für verschiedene Lemmata zu kategorisieren, Kollokationen dazu zu berechnen und Frequenzverteilungen in verschiedenen zeitlich definierten Teilkorpora auszugeben. Um die Relevanz von Frequenzunterschieden zu berechnen verwendet sie statistische Signifikanztests.

Als Studentin der Computerlinguistik kann Angela Fahrni in **„Regelmässigkeiten in Kundenrezensionen auf Amazon“** (Fahrni, 2008) auf avanciertere Methoden zurückgreifen. Mit der Hilfe von eigenen Perl-Scripten erstellt sie sich ein Korpus von 39.063 Kundenrezensionen von `amazon.de`, das sie im XML-Format, angereichert mit den verfügbaren Metadaten, speichert. Zudem setzt sie den „TreeTagger“ (Schmid, 1994) ein, um die Daten mit Part-of-Speech-Tags zu taggen. Den Tagger ergänzt sie um eigene Wortklassen, um Emoticons zu annotieren. Für die Analyse verwendet sie die Konkordanz- und Recherchesoftware „XAIRA“ (Oxford University Computing Services 2011, vgl. Abbildung 2), die XML-Dokumente verarbeiten kann. Zudem benutzt sie „gCLUTO – Graphical Clustering Toolkit“ (Rasmussen/Karypis, 2004), um zu berechnen, welche sprachlichen Marker positive und negative Rezensionen gut voraussagen können.

Das **„Wörterbuch der Krise“** (Baumgärtner u. a., 2010)[8] ist ein kollaboratives Werk von den Studierenden Verena Baumgärtner, Sascha Braun, Barbara Katharina Dietz, Verena Keite, Maximilian Nowroth, Frederic Wagner und Johannes Wolf.[9] Ziel war, mit korpuslinguistischen Mitteln drei unterschiedliche Krisen-Diskurse lexikographisch anzugehen. Grundlage bilden Recherchen in öffentlichen Korpora (DeReKo IDS, o. J.; DWDS, o. J.), aber auch eigens zusammengestellte Zeitungskorpora, die vor allem mit Kollokationsanalysen bearbeitet wurden.

Eine besondere Datengrundlage verwendet Madeleine Ehrensperger für ihre Arbeit **„Geschlechts- und Altersspezifisches Sprachverhalten“** (Ehrensperger, 2006): Mittels eines Fragebogens mit politischen/gesellschaftlichen Einstellungsfragen erhält sie von 60 Versuchspersonen beiderlei Geschlechts und unterschiedlichen Alters schriftlich geäußerte Statements. Diese untersucht sie hinsichtlich der Verwendung der linguistischen Parameter Satzlänge, Ich-Aussagen, Satzklammern, Abkürzungen und Ausrufe-

[7] „AntConc“ kann keine annotierten Korpora verarbeiten. Im Fall der beschriebenen Arbeit war es aber möglich, die gewünschten Redebeiträge aus dem annotierten Korpus zu extrahieren und als nicht-annotierte Textdateien in AntConc zu laden.

[8] Online erreichbar über `http://www.bubenhofer.com/Krise/` (23. März 2011).

[9] Mitbetreuerin dieser Arbeiten war Stefaniya Ptashnyk.

und Fragezeichen, um herauszufinden, ob bezüglich dieser Parameter ein geschlechtsspezifisches Schreiben beobachtbar ist. Sie nutzt keinerlei technische Hilfsmittel, obwohl das nahegelegen hätte.

2.3 Hoffnungen und Enttäuschungen

Als Betreuer der oben kurz dargestellten Arbeiten machte ich die Erfahrung, dass die Studierenden es generell schätzten, empirisch arbeiten zu können. Ebenso attraktiv scheint es zu sein, dass diese Arbeiten sehr anwendungsbezogen und nicht primär eine theoretische Auseinandersetzung sind. Die Korpuslinguistik kann zudem bis zu einem gewissen Grad das Bedürfnis stillen, mehr oder weniger klar definierte Methoden für die Analyse von Sprachdaten anwenden zu können.

Doch diesen positiven Aspekten müssen auch Enttäuschungen der Studierenden gegenüber gestellt werden. In erster Linie treten diese ein, wenn der große technische Aufwand deutlich wird. Empirisches Arbeiten ist aufwändig und wenn gleichzeitig noch die technischen Fertigkeiten erlangt werden müssen, um die Werkzeuge anwenden zu können, kann dies zu Frustgefühlen führen.

Bei einigen Arbeiten stellt sich zudem das Problem der Operationalisierung der Hypothesen, die vor dem Hintergrund linguistischer Theoriebildung, gerade im Bereich der Diskurs- und Kulturanalyse, sehr schwierig ist. Damit verbunden ist dann die Enttäuschung über die (vermeintlich) beschränkte Aussagekraft von korpuslinguistischen Analyseresultaten. N-Gramm-Analysen oder Kollokationsberechnungen führen rasch zu großen Datenmengen, die gesichtet und kategorisiert werden müssen; die Analyse auf Knopfdruck ist eine Utopie.

3 Nutzerprofil Online-Kurs Korpuslinguistik

Meine Website „Einführung in die Korpuslinguistik. Praktische Grundlagen und Werkzeuge“ (Bubenhofer, 2006-2011) entstand im Rahmen meiner gleichnamigen Veranstaltung an der Universität Zürich (vgl. Abbildung 3). Seit 2006 gab es immer wieder kleinere Aktualisierungen und Erweiterungen, so dass das Angebot inzwischen die in Tabelle 2 dargestellten Themen umfasst.

Gemäß Zugriffsstatistik gab es im Jahr 2010 14.158 Besuche mit 43.319 Seitenaufrufen, wobei 34% über direkte Zugriffe[10], 21% über Verweise und 45% über Suchmaschinen zustande gekommen sind. Aus der Analyse der Verweise lässt sich schließen, dass das Angebot auch an einigen Universitäten in der Lehre benutzt wird.

Die Abbildungen 4 und 5 zeigen die am häufigsten und am längsten aufgerufenen Seiten der Einführung. Nicht weiter überraschend liegen grundlegende Themen wie „Einführung“, „Definitionen“ oder „Korpustypen“ ganz oben. Auch die Ausführungen zu „DeReKo/COSMAS II“ werden stark nachgefragt.

[10] Die Adresse wurde also direkt in den Browser eingegeben oder über ein gespeichertes Lesezeichen abgerufen.

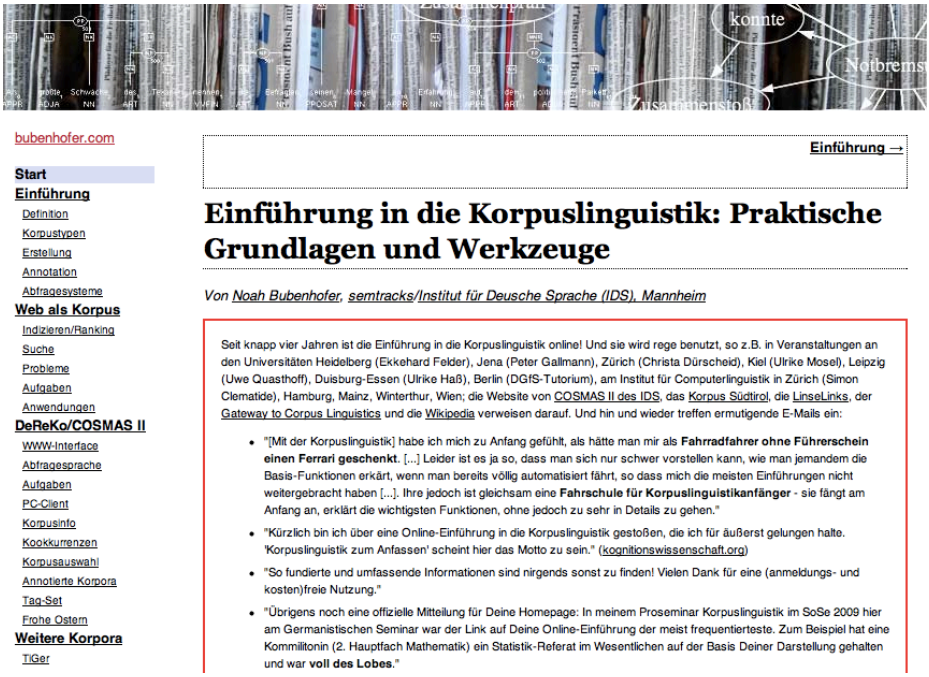


Abbildung 3: Die Startseite der „Einführung in die Korpuslinguistik“ (Bubenhofer, 2006-2011), erreichbar unter `www.bubenhofer.com/korpuslinguistik/`.

1. **Einführung:** Definition Korpuslinguistik, Korpustypen, Erstellung von Korpora, Annotation, Abfragesysteme
2. **Web als Korpus:** Funktionsweise von Suchmaschinen, Suchmöglichkeiten, Probleme, Anwendungen
3. **DeReKo/COSMAS II:** Informationen zur Funktions- und Verwendungsweise des DeReKo IDS (o. J.)
4. **Weitere Korpora:** TiGer (Lezius, 2002; Brants u. a., 2002), DWDS (o. J.), Wortschatz Leipzig (o. J.), Archiv für gesprochenes Deutsch des Instituts für Deutsche Sprache (Haas/Wagener, 1992)
5. **Eigenes Korpus:** Daten beschaffen, aufbereiten, analysieren, „AntConc“ (Anthony, 2010), „kfNgram“ (Fletcher, 2010)
6. **Corpus Workbench:** Die Arbeit mit der Open Corpus Workbench (CWB, 2011)
7. **Datenbank Filemaker:** Einführung, Datenbank erstellen, durchsuchen, Daten importieren, CSV-Formatierung, exportieren
8. **Anwendungen:** Forschungsprozess, Semantik, Argumentationsmuster, Diskursanalyse
9. **Statistik:** Einführung, Signifikanztests
10. **Visualisierung:** Möglichkeiten, GraphViz (Ellson u. a., 2011), Beispiele
11. **Anhang:** Software, Unix-Befehle, Reguläre Ausdrücke, Literatur, Lexikon, Impressum

Tabelle 2: Inhalt der „Einführung in die Korpuslinguistik“ (Bubenhofer, 2006-2011).

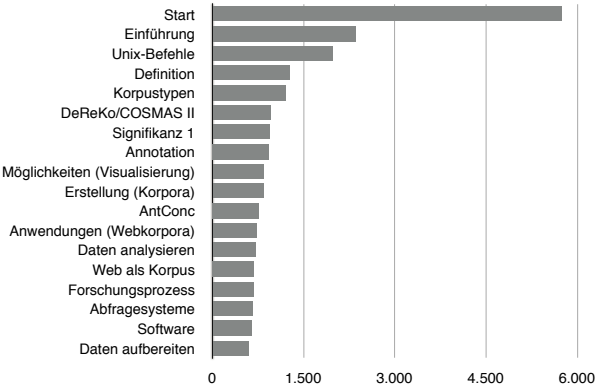


Abbildung 4: Die im Jahr 2010 am häufigsten aufgerufenen Seiten der „Einführung in die Korpuslinguistik“ (Anzahl Seitenaufrufe).

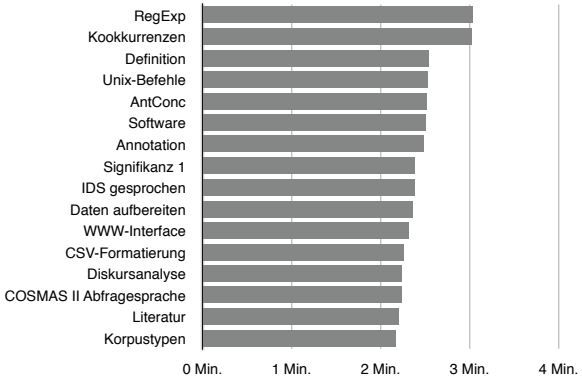


Abbildung 5: Die im Jahr 2010 am längsten aufgerufenen Seiten der „Einführung in die Korpuslinguistik“ (durchschnittliche Anzahl Minuten).

Die Seiten „Reguläre Ausdrücke“ und „Kookkurrenzen“ führen zu den höchsten Verweildauern. Ebenso weitere textlastige Ausführungen zu Software, Annotation, Statistik etc. In beiden Abbildungen ist zudem sichtbar, dass die Seite zu den Unix-Befehlen stark nachgefragt wird. Es handelt sich hierbei um eine Sammlung von grundlegenden Unix-Befehlen, die für korpuslinguistische Fragestellungen nützlich sind. Allerdings muss hierbei betont werden, dass sowohl diese Seite als auch jene zu den regulären Ausdrücken wohl nicht nur von Besuchern nachgefragt werden, die ein korpuslinguistisches Interesse mitbringen, sondern vor allem über Suchmaschinen Besucher anziehen, die aus anderen Gründen danach suchen.

Da 45% der Besuche über Suchmaschinen zugeführt werden, analysierte ich zusätzlich die Suchbegriffe, mit denen die Besucher auf die Seite gelangen (Tabelle 3). Hierbei muss natürlich beachtet werden, dass nur Suchbegriffe angezeigt werden, mit denen man auf der Website fündig werden kann. Die Liste kann also nicht aufzeigen, welche Themen im Angebot fehlen, sondern nur Hinweise darauf geben, welche Aspekte der bestehenden Inhalte auf besonders großes Interesse stoßen.

Es wird ersichtlich, dass oft nach „Themen“ oder „Anwendungsgebieten“ der Korpuslinguistik gesucht wird. Es scheint also wichtig zu sein, Anregungen für korpuslinguistisches Arbeiten zu geben. Weiter gibt es diverse Suchbegriffe, die die Korpuserstellung zum Thema haben oder gezielt Informationen zu bestimmten bestehenden Korpora suchen.

Natürlich sind auch Hinweise auf und Anleitungen von Programmen für korpuslinguistische Arbeiten gefragt. Solche Anleitungen scheinen auch dann ein Bedürfnis zu sein, wenn der Softwareanbieter eigentlich eine Dokumentation anbietet, wie z. B. bei „AntConc“, wo die offizielle Anleitung allerdings nur auf Englisch zur Verfügung steht und relativ knapp gehalten ist. Auch der Erklärungsbedarf von Methoden wie Kookkurrenzanalyse oder statistische Signifikanztests scheint groß zu sein.

Zusammenfassend kann gesagt werden, dass die Benutzer der Website einerseits Werkzeuge suchen für die Recherche in Korpora, das Erstellen und Verwalten von Korpora, die statistische Analyse und die Annotation, andererseits aber auch nach Hilfen und Anleitungen zur Bedienung dieser Werkzeuge. Und natürlich werden Inspirationsquellen gewünscht, die korpuslinguistisches Arbeiten zeigen.

4 Fazit: Möglichkeiten und Wünsche

Ich habe versucht zu zeigen, welche Inhalte eine Einführung in die Korpuslinguistik in der linguistischen Lehre umfassen kann und was für Projekte realistisch sind, in denen die Studierenden das Gelernte anwenden können. Zudem hat die Analyse des Web-Angebots meiner Einführung in die Korpuslinguistik aufgezeigt, welche Themen besonders beliebt sind.

Aus diesen Erfahrungen heraus kann ich nun als Lehrperson einige Wünsche formulieren, die in erster Linie das Angebot an korpuslinguistischem Werkzeug betrifft. Für die Lehre wäre es ein Desiderat, über Software-Module für unterschiedliche Anwendungen zu verfügen:

Inhalte	
themen korpuslinguistik	130
anwendungsgebiete korpuslinguistik	26
diskursanalyse	21
korpuslinguistik diskursanaylse	11
probleme der korpuslinguistik	5
Korpuserstellung/Korpora	
korpus definition	36
korpuslinguistik tageszeitungen	43
korpus erstellen	34
daten aufbereiten	5
erstellung ein korpus	5
wie erstelle ich einen korpus	5
textkorpus erstellen download	12
filemaker datenbank erstellen	9
korpuslinguistik copyright	25
deutschsprachige korpora	6
korpustyp	6
baumbanken	5
tiger corpus	41
cosmas ii	26
ids korpus	9
funktionen cosmas	6
dereko	5
dwds	15
lexis nexis korpus	7
Annotation	
annotation korpuslinguistik	23
annotierte korpora	12
korpuslinguistik tagging	5
pos tagger online	7
tagset	54
korpulinguistik tag sets	5
dependenz parser	5

Software	
konkordanzprogramm	33
konkordanzprogramm download	5
concordance-programme zur analyse von korpora	5
korpuslinguistik tools	7
simple concordance program	12
korpuslinguistik software	21
textdatei importieren per script filemaker	13
antconc	277
antconc anleitung	7
antconc regex	6
cluster antconc	5
t-score antconc	5
graphviz	12
graphviz beispiele	7
graphviz dot	7
graphviz gui	5
kfngram	12
filemaker	8
tigersearch	14
corpus workbench windows	7
treetagger betriebssystem	15
regexp	5
reguläre ausdrücke antconc	5
software berechnung signifikanz	5
Statistik	
kookkurrenzen	45
kookkurrenzanalyse	15
kookkurrenzprofil	14
log likelihood test	30
log-likelihood	10
llr wert	15
log likelihood tabelle	6
chi quadrat test signifikant	5
signifikanz	8
signifikanztest excel	8
kontingenztabelle signifikanz	13
darstellungsoptionen konkordanz korpuslinguistik	8

Tabelle 3: Suchbegriffe von Suchmaschinen, die zur „Einführung in die Korpuslinguistik“ führen; grob geordnet nach thematischen Bereichen (zweite Spalte: Anzahl Suchende).

- Korpuserstellung (Textaufbereitung, Web-Download etc.), Verwaltung, Annotation, Analyse und Ergebnisdarstellung.
- Die Softwaremodule sollten über einheitliche Schnittstellen verfügen, so dass man die Daten und Ergebnisse leicht mit unterschiedlichen Modulen weiterverarbeiten kann.
- Die Module sollten möglichst plattformunabhängig sein, denn die verfügbare Infrastruktur, sei es in PC-Pools an der Universität, seien es die privaten Rechner der Studierenden, ist bezüglich Betriebssystemen sehr heterogen.[11]
- Wichtig ist die einfache Bedienbarkeit über eine leicht verständliche grafische Benutzeroberfläche. Nur so können Studierende angesprochen werden, die keine große Affinität zu Computern haben.

Zum wichtigsten Ziel von korpuslinguistischen Kursen zähle ich, die Studierenden überhaupt dazu zu motivieren, korpuslinguistisch, also empirisch zu arbeiten. Dabei ist es wichtig, die Angst vor den technischen Hürden zu nehmen und die Studierenden dazu zu ermutigen, auch Methoden zu verwenden, die auf den ersten Blick kompliziert wirken. Dazu gehört z. B. auch der Einsatz von statistischen Methoden, etwa zur Überprüfung von Signifikanzunterschieden.

Dabei bewegt man sich unweigerlich auf dem schmalen Grat des Realistischen und potenziell Möglichen: Die State-of-the-Art der Korpuslinguistik ist immer fortgeschrittener als das, was in der linguistischen Lehre tatsächlich gemacht werden kann. Mit anschaulichen Beispielen, die man auf einfach bearbeitbare Teilaufgaben herunterbricht, können die Studierenden jedoch motiviert werden, sich an den Stand der Kunst heranzutasten.

Doch bei allen technischen Hürden schien es mir immer wichtig, zunächst von linguistisch fundierten Hypothesen auszugehen, diese zu operationalisieren und erst dann zu prüfen, ob das Vorhaben technisch umgesetzt werden kann. Es wäre schade, wenn man sich schon gleich zu Beginn vom technischen Aufwand abschrecken ließe.

Literatur

Anthony, Laurence (2010): *AntConc 3.2.1* <`http://www.antlab.sci.waseda.ac.jp/`>.

Baertschi, Sara (2010): Der Berg ruft. Sprachgebrauchsmuster von 1920-1945 in der Literatur des Schweizer Alpen-Clubs [Unveröffentlichte Lizentiatsarbeit, Deutsches Seminar, Universität Zürich].

Baumgärtner, Verena/Braun, Sascha/Dietz, Barbara Katharina/Keite, Verena/Nowroth, Maximilian/Wagner, Frederic/Wolf, Johannes (2010): *Wörterbuch der Krise* <`http://www.bubenhofer.com/Krise/`> [betreut von Stefaniya Ptashnyk und Noah Bubenhofer].

[11] Zwar sind Windows-Rechner sehr verbreitet, besonders an Schweizer Universitäten sind jedoch auch viele Mac-Systeme in Verwendung und in Computerlinguistik-Kontexten sind oft Linux-Systeme im Einsatz.

Brants, Sabine/Dipper, Stefanie/Hansen, Silvia/Lezius, Wolfgang/Smith, George (2002): The TIGER Treebank. In: *Proceedings of the Workshop on Treebanks and Linguistic Theories*, Sozopol.

Bubenhofer, Noah (2006-2011): *Einführung in die Korpuslinguistik: Praktische Grundlagen und Werkzeuge*. Elektronische Ressource <`http://www.bubenhofer.com/korpuslinguistik/`>.

Bubenhofer, Noah (2009): *Sprachgebrauchsmuster. Korpuslinguistik als Methode der Diskurs- und Kulturanalyse*. Berlin, New York: de Gruyter (Sprache und Wissen; 4).

Bubenhofer, Noah/Dussa, Tobias/Ebling, Sarah/Klimke, Martin/Rothenhäusler, Klaus/Scharloth, Joachim/Tamekue, Suarès/Vola, Saskia/Forschergruppe semtracks (2009): „So etwas wie eine Botschaft". Korpuslinguistische Analysen der Bundestagswahl 2009. In: *Sprachreport* 4, S. 2–10.

Bubenhofer, Noah/Klimke, Martin/Scharloth, Joachim (2008): *political tracker – U.S. Presidential Campaign '08: A Semantic Matrix Analysis*. Elektronische Ressource <`http://semtracks.com/politicaltracker/`>.

Casana, Verena (2009): Homosexualität. Analyse der Paradigmengruppe *homosexuell – gleichgeschlechtlich* anhand der taz [Unveröffentlichte Hausarbeit, Deutsches Seminar, Universität Zürich].

CWB (2011): *The IMS Open Corpus Workbench (CWB)* <`http://cwb.sourceforge.net/`>.

DeReKo IDS (o. J.): *Das Deutsche Referenzkorpus DeReKo*. Elektronische Ressource <`http://www.ids-mannheim.de/kl/projekte/korpora/`>.

DWDS (o. J.): *Digitales Wörterbuch der deutschen Sprache des 20. Jahrhunderts*. Elektronische Ressource <`http://www.dwds.de`>.

Ehrensperger, Madeleine (2006): Geschlechts- und Altersspezifisches Sprachverhalten [Unveröffentlichte Hausarbeit, Deutsches Seminar, Universität Zürich].

Ellson, John/Gansner, Emden/Hu, Yifan/Bilgin, Arif (2011): *Graphviz – Graph Visualization Software* <`http://www.graphviz.org`>.

Fahrni, Angela (2008): Regelmässigkeiten in Kundenrezensionen auf Amazon [Unveröffentlichte Hausarbeit, Deutsches Seminar, Universität Zürich].

Fletcher, William H. (2010): *kfNgram: Information and Help* <`http://www.kwicfinder.com/kfNgram/`>.

Google Books Ngram Viewer (o. J.): *Google Books Ngram Viewer*. Elektronische Ressource <`http://ngrams.googlelabs.com/`>.

Haas, Walter/Wagener, Peter (Hgg.) (1992): *Gesamtkatalog der Tonaufnahmen des Deutschen Spracharchivs. Erarbeitet von Mitarbeiterinnen und Mitarbeitern des Instituts für Deutsche Sprache*. Tübingen: Niemeyer (Phonai; 38/39).

Lemnitzer, Lothar (2011): *Die Wortwarte* <`http://www.wortwarte.de`>.

Lemnitzer, Lothar/Zinsmeister, Heike (2006): *Korpuslinguistik. Eine Einführung*. Tübingen: Narr.

Lezius, Wolfgang (2002): *Ein Suchwerkzeug für syntaktisch annotierte Textkorpora*. Phil. Diss. University of Stuttgart, Stuttgart <http://www.ims.uni-stuttgart.de/projekte/corplex/paper/lezius/diss/disslezius.pdf>.

Lieberman, Erez/Michel, Jean-Baptiste/Jackson, Joe/Tang, Tina/Nowak, Martin A. (2007): Quantifying the evolutionary dynamics of language. In: *Nature* 449, S. 713–716.

Matic, Igor (2009): Konzeptuelle Metaphern der Wirtschaftskrise in der NZZ am Sonntag [Unveröffentlichte Hausarbeit, Deutsches Seminar, Universität Zürich].

Michel, Jean-Baptiste/Shen, Yuan Kui/Aiden, Aviva Presser/Veres, Adrian/Gray, Matthew K./Team, The Google Books/Pickett, Joseph P./Hoiberg, Dale/Clancy, Dan/Norvig, Peter/Orwant, Jon/Pinker, Steven/Nowak, Martin A./Aiden, Erez Lieberman (2011): Quantitative Analysis of Culture Using Millions of Digitized Books. In: *Science* 331, H. 6014, S. 176–182 <http://www.sciencemag.org/content/331/6014/176.abstract>.

Oxford University Computing Services (2011): *All About Xaira* <http://www.oucs.ox.ac.uk/rts/xaira/>.

Papst, David (2005): *klein* und *winzig*. Eine Korpusuntersuchung zur Synonymie [Unveröffentlichte Hausarbeit, Deutsches Seminar, Universität Zürich].

Rasmussen, Matt/Karypis, George (2004): gCLUTO: An Interactive Clustering, Visualization, and Analysis System. *Techn. Ber. 04-021*, University of Minnesota Technical Report <http://glaros.dtc.umn.edu/gkhome/cluto/gcluto/overview>.

Rauber, Paul (2009): Intellektuelle im Diskurs. Zwischen Hybris und Machtkritik [Unveröffentlichte Hausarbeit, Deutsches Seminar, Universität Zürich].

Scherer, Carmen (2006): *Korpuslinguistik*. Heidelberg: Winter (Kurze Einführungen in die Germanistische Linguistik; 2).

Schmid, Helmut (1994): *Probabilistic Part-of-Speech Tagging Using Decision Trees* <http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger1.pdf>.

Text+Berg-Korpus (2011): *Text+Berg-Korpus (Release 145)*. XML-Format [Digitale Edition des Jahrbuch des SAC 1864-1923 und Die Alpen 1925-2009].

Weibel, Tamara (2009): *Mieterinnen* oder *Mieter – Schweizervolk* oder *Ausländer*? Parteispezifische Personen- und Personengruppenbezeichnungen der SP und SVP im Schweizer Parlament [Unveröffentlichte Hausarbeit, Deutsches Seminar, Universität Zürich].

Wortschatz Leipzig (o. J.): *Deutscher Wortschatz Universität Leipzig*. Elektronische Ressource <http://wortschatz.uni-leipzig.de>.

Autorenindex

Maja Bärenfänger
Justus-Liebig-Universität Gießen
maja.baerenfaenger@germanistik.uni-giessen.de

Michael Beißwenger
Technische Universität Dortmund
michael.beisswenger@uni-dortmund.de

Frank Binder
Justus-Liebig-Universität Gießen
frank.binder@germanistik.uni-giessen.de

Noah Bubenhofer
Institut für Deutsche Sprache, Mannheim
bubenhofer@ids-mannheim.de

Katja Bülow
The Open University
k.buelow@open.ac.uk

Stefanie Dipper
Ruhr-Universität Bochum
dipper@linguistics.rub.de

Rüdiger Gleim
Goethe-Universität Frankfurt am Main
gleim@em.uni-frankfurt.de

Debra Haley
The Open University
d.t.haley@open.ac.uk

Masatoshi Ishikawa
Tokyo Seitoku University
ishikawa@tsu.ac.jp

Bernhard Jussen
Goethe-Universität Frankfurt am Main
jussen@em.uni-frankfurt.de

Kei-ichi Kaneko
Tokyo University of Agriculture and Technology
k1kaneko@cc.tuat.ac.jp

Henning Lobin
Justus-Liebig-Universität Gießen
Henning.Lobin@uni-giessen.de

Harald Lüngen
Institut für Deutsche Sprache, Mannheim
luengen@ids-mannheim.de

Alexander Mehler
Goethe-Universität Frankfurt am Main
mehler@em.uni-frankfurt.de

Haruko Miyakoda
Tsuda College
miyakoda@tsuda.ac.jp

Silke Schwandt
Goethe-Universität Frankfurt am Main
schwandt@em.uni-frankfurt.de

Angelika Storrer
Technische Universität Dortmund
angelika.storrer@uni-dortmund.de

Maik Stührenberg
Universität Bielefeld
maik.stuehrenberg@uni-bielefeld.de

Fridolin Wild
The Open University
f.wild@open.ac.uk

Sabrina Wilske
Universität des Saarlandes
sw@coli.uni-saarland.de

Magdalena Wolska
Universität des Saarlandes
magda@coli.uni-saarland.de

Heike Zinsmeister
Universität Konstanz
Heike.Zinsmeister@uni-konstanz.de