Creation and Analysis of a Reading Comprehension Exercise Corpus: Towards Evaluating Meaning in Context

Niels Ott, Ramon Ziai and Detmar Meurers

Abstract

We discuss the collection and analysis of a cross-sectional and longitudinal learner corpus consisting of answers to reading comprehension questions written by adult second language learners of German. We motivate the need for such task-based learner corpora and identify the properties which make reading comprehension exercises a particularly interesting task.

In terms of the creation of the corpus, we introduce the web-based WELCOME tool we developed to support the decentralized data collection and annotation of the richly structured corpus in real-life language teaching programs. On the analysis side, we investigate the binary and the complex content-assessment classification scheme used by the annotators and the inter-annotator agreement obtained for the current corpus snapshot, at the halfway point of our four-year effort. We present results showing that for such task-based corpora, meaning assessment can be performed with reasonable agreement and we discuss several sources of disagreement.

1 Introduction

This paper discusses the creation and analysis of a corpus which is motivated by a general research question: How can the meaning of sentences and text fragments be analyzed and compared in realistic situations, where the language used is not necessarily well-formed and there are differences in situative and world knowledge? To address this research question, one needs to determine which linguistic representations can be robustly identified as empirical basis of the computational

approximation of meaning. At the same time, it is important to make concrete and investigate the role of the context of the sentences to be analyzed. It is in pursuit of those issues that we started to collect data of authentic language in context. We are focusing on reading comprehension exercises which appear particularly well-suited given that they are a common task in real-life language teaching and such exercises include an explicit linguistic context in the form of the questions and the text these are about.

The notion of a task and language use in context plays an important role in foreign language teaching and learning (Ellis, 2003). Correspondingly, a representation of the learner's ability to use language in context and perform tasks using appropriate strategies has also been argued to be crucial for interpreting learner language and for informing learner modeling in Intelligent Tutoring Systems (Amaral and Meurers, 2008). Most learner corpora, however, consist of learner essays with minimal requirements on meaning and form (Granger, 2008). Borrowing the terminology of Bachman and Palmer (1996), such learner essays are *indirect responses*, primarily encoding individual knowledge of the learners.

Fitzpatrick and Seegmiller (2004) show that for such corpora, it can be very difficult to interpret the learner data. For the Montclair corpus they fail to reach sufficiently high inter-annotator agreement levels for annotating target hypotheses, an essential component of error annotation (Hirschmann et al., 2007, sec 2.3.1). To the best of our knowledge, to date there is no inter-annotator agreement study of any error annotation scheme establishing which distinctions can reliably be annotated for learner corpora.

To illustrate the difficulty of interpreting learner language without an explicit task context, consider the learner sentences in (1) from the Hiroshima English Learners' Corpus (Miura, 1998).

(1) a. I didn't know
    b. I don't know his lives.
    c. I know where he lives.
    d. I know he lived

Based solely on these learner sentences, it seems that the target sentences are easy to determine, especially given that target forms diverging from what the learner wrote typically are only considered

when the learner sentences appear ungrammatical. Yet, when we take into account that these learner sentences were produced in a translation task, for which the teacher provided the target translation shown in (2), it becomes apparent how important this task information is for interpreting the learner responses in (1).

(2) I don't know where he lives.

To support reliable interpretation of the form and meaning of learner data, it clearly is important to collect such data with explicit task contexts. For the case of the reading comprehension questions we are focusing on, a simple way of expressing this is that it is easier to infer what a learner wanted to say if one knows the text they are answering questions about.

Looking at the nature of the interpretation and annotation of learner language more broadly, it is closely related to that found in educational assessment of learner performance. For the latter, Mislevy (2006) highlights that the "interpretation of [a student's] actions rather than the actions themselves constitute data in an assessment argument" and points to three aspects to consider: "(i) aspects of the situation in which the person is acting, (ii) aspects of the person's action in the situation, and (iii) additional information about the person's history or relationship to the observational situation." For interpreting language learner data, we thus need to consider (i) the task for which a learner is producing the language, (ii) the language produced, and (iii) the personal and the interaction history of the learner. Note that while the task aspect has not yet played much of a role in learner corpus research, learner variables are already being standardly collected in the metadata of learner corpora to support inferences based on the first language and other learner properties.

Collecting learner data with explicit task contexts also make it possible to take task-specific learner strategies into account. For example, learners do not always aim at expressing a particular meaning and instead may prioritize writing a well-formed answer. Indeed, in our work with learners of English (Bailey, 2008; Bailey and Meurers, 2008), we found that learners often lift material from texts or use chunks familiar to the learner instead of trying to formulate new sentences expressing the appropriate meaning. This strategy was more frequently used by the less proficient learners so

that as a result the less proficient learners made fewer form errors in their responses – an observation which clearly could not be explained without taking the task and the task strategies of the learners into account.

In light of these general issues speaking for the collection of task-based learner corpora, for our research on the automated comparison of meaning we explore the creation and annotation of a corpus in which the learner language produced is explicitly contextualized and directly related to the prompt and input that is provided by a given reading comprehension task.

2 The Corpus of Reading Comprehension Exercises in German (CREG)

Our corpus is being collected as part of an ongoing four-year project in collaboration with two large German programs in the US, at the University of Kansas (Prof. Nina Vyatkina) and at The Ohio State University (Prof. Kathryn Corl). The regular reading comprehension exercises that are used in German classes are collected and evaluated by assistants with previous experience in German teaching in those programs who were hired for this project. Data is collected at beginner, intermediate and advanced levels of instruction during the entire project. Both universities offer several courses at the same level simultaneously and courses at different levels are held each semester, which will result in the first large learner corpus including an explicit task context.

We intentionally focus on the relatively homogeneous foreign language learner populations at these Midwestern universities, where the students' exposure to German is mostly through the classroom setting. Second language learners of German studying at a German university would have a wide range of first language backgrounds and are exposed to the influence of everyday life interactions in German. In addition to the text and task data, we also collect metadata on the learners. This includes background information such as age, gender, previous exposure to German, other foreign languages learned, and time spent in a German-speaking country. Since the corpus collection effort lasts four years and may track individual learners over this time period, some of the metadata is bound to change. To account for this, updated versions of each student's metadata are

collected at the beginning of each term, yielding a connected history of individual student records (using anonymous identifiers). These student records mark points in time that can be related to the actual learner performances in the corpus. In addition, all submissions from the students are equipped with date stamps. Hence, the student metadata records together with the submissions make it possible to track the changes in a learner's production longitudinally.

The meaning of each learner answer is assessed by two independent annotators with German teaching experience. While in administrative terms they were hired as research assistants for the project, in this paper we refer to them as teachers to underline the fact that they are essentially grading the answers to standard reading comprehension exercises rather than performing a more artificial, research-defined linguistic annotation.

Meaning assessment is carried out as a binary classification (*appropriate* vs. *inappropriate* meaning), with form errors (spelling, grammar) not being taken into account in the assessment. To gain insights into the grader observations leading to the assessment, we additionally asked for more detailed meaning assessment categories encoding the nature of the divergence from the target answers specified by the teachers. Following Bailey and Meurers (2008), we distinguish *missing concept*, *extra concept*, *blend* (missing concept and extra material), and finally *non-answer* for answers which are unrelated to the question under discussion.

The resulting highly structured corpus is represented in a relational database which supports references between the different kinds of data. The overall corpus layout is visualized in Figure 1.

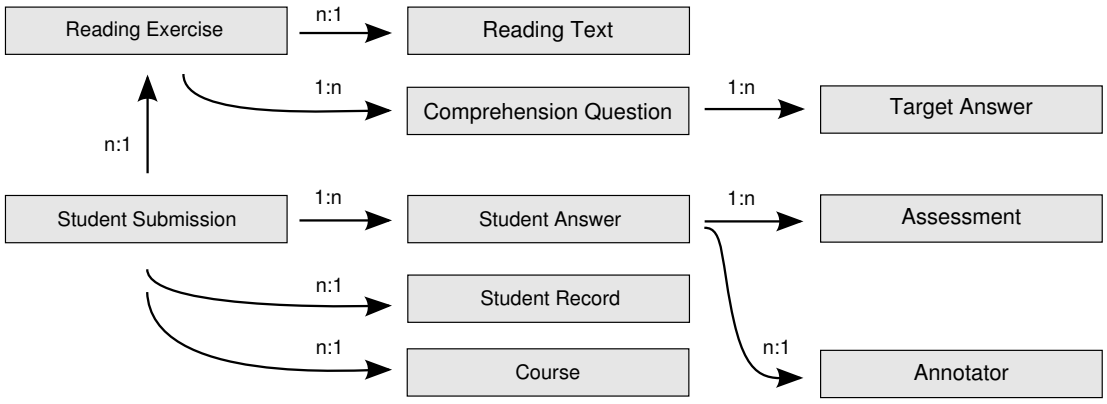

Figure 1: CREG Corpus Layout

The reading texts are linked to the specific exercises that make use of them, along with the questions that pertain to the text in the specific exercise. In addition to the textual information entered, the original exercise sheet is also saved as a file together with each exercise. Each exercise can include multiple questions, where each question can again have multiple target answers. Student submissions are stored for each completed exercise sheet. They consist of the student answers themselves as well as references to the other relevant parts of a submission, the exercise and the student. The student answers contain references to their meaning assessment, as well as to the target answer they most closely resemble.

While we originally designed the corpus with a particular research question in mind, the resulting longitudinal learner corpus with its task and learner information should be rich enough to support a range of research perspectives, from second language acquisition research into the specifics of learner language and interlanguage development, via theoretical linguistic research into information structure, to computational linguistic analysis of answers to reading comprehension questions and the use of such analysis in ICALL systems.

3 Corpus Collection and the WELCOME tool

Language teachers are the domain experts on reading comprehension exercises and they regularly assign exercises to language learners and evaluate them. Yet, they are not typically experts in computer use, corpus collection, and annotation tools and cannot simply be asked to use data entry methods requiring advanced computer skills, such as editing XML data or managing files with revision control systems. Therefore, it was necessary to provide a tool that supports the language teachers in collecting the structured learner corpus, i.e., the exercises, the learner data, and the meaning assessment for that data. We wanted the tool to support data entry of a richly structured corpus and metadata by several teachers at different sites in a distributed manner and to store the corpus and metadata in a central state-of-the-art corpus repository.

To address the desiderata, we developed the WEb-based Learner Corpus MachinE (WELCOME,

http://purl.org/icall/welcome), a web-based application that at the same time enforces the corpus structure and fits in with the workflow and task conceptualization of language teachers. WELCOME behaves similar to a desktop application, but it does not need to be installed on the teachers' computers, requiring only a recent web browser and an Internet connection to run. All data are stored centrally on our servers in a relational database system, for which regular revision control snapshots and backups are ensured. The design of the database not only reflects the structure discussed in section 2, it also enforces relations between corpus elements to be established by constraints, ensuring and maintaining data consistency. Since its first presentation at Linguistic Evidence 2010 (Meurers, Ott, and Ziai, 2010), WELCOME has been constantly improved and adapted to the users needs.

The general procedure of collecting data using WELCOME is the following: First, teachers create reading exercises consisting of a title, task description, reading text and a number of questions including target answers. The frontend for this step is depicted in Figure 2. Second, the student metadata is entered by the teachers (once per term). Third, the paper-based submissions containing the answers to the reading comprehension questions are collected from the students. These answer sheets are then scanned in and uploaded, given that transcribing from handwriting is an act of interpretation requiring explicit documentation. Subsequently, two teachers work independently on the answer sheets: They transcribe the student answers and provide the binary and the detailed meaning assessment, as illustrated in Figure 3. Since some exercises are typed and submitted online by the students, the exercise editor is equipped with a little checkbox to mark such cases. In this case, the student answers can simply be copy-pasted from some online learning platform and thus do not require a second transcription.

To provide access to the corpus, WELCOME includes a function exporting the current database in a dedicated XML format. This format can be used for distributing the corpus data and serves as an input format for several of the tools developed in our research project (Meurers, Ziai, Ott, and Bailey, 2011). WELCOME is not specific to a particular language and is freely available under a Creative Commons by-nc-sa license for reuse by other corpus collection efforts.

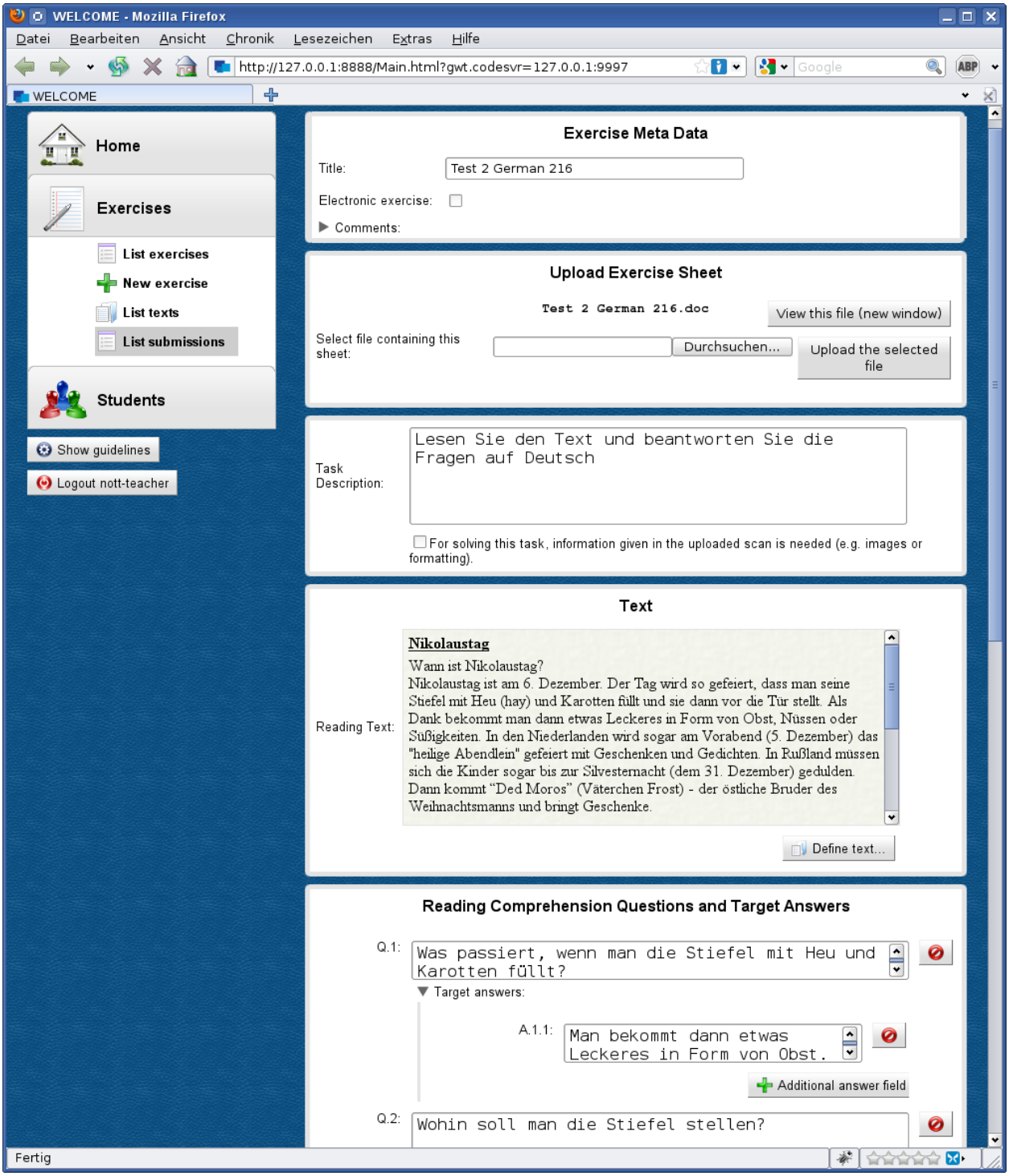

Figure 2: Creating an exercise template using WELCOME

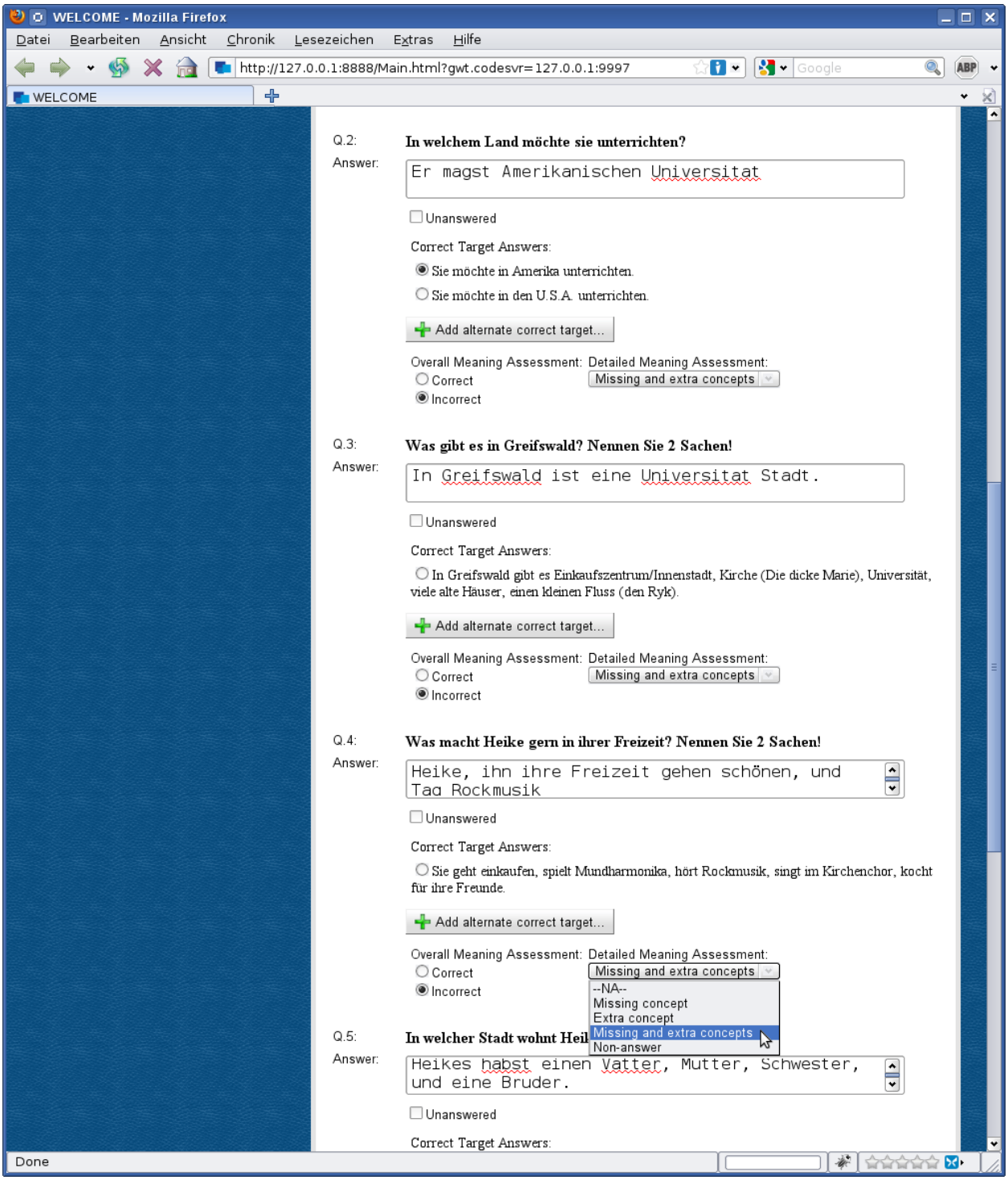

Figure 3: Meaning assessment in WELCOME

4 Inter-Annotator Agreement Analysis for Meaning Assessment

4.1 Quantitative Results

In this section we investigate a snapshot of the corpus created from both the KU and the OSU data on May 25, 2011 via WELCOME's XML exporter. This snapshot was converted into tables with student answers, question IDs and meaning assessments that can be examined using the statistical software R (R Development Core Team, 2009). In both universities, the annotators hired for this project are experienced in teaching German and grading reading comprehension exercises. Both at KU and at OSU one of the annotators left the team and was replaced by another annotator. The third annotator worked only on those answers which the original annotator did not yet deal with. As a result there are three annotators for each subcorpus, but each student answer was annotated by exactly two annotators. We call these two annotators A1 and A2, with the latter being the combination of the original annotator and the replacement. Student answers that so far have been annotated only by one person were excluded from this study.

Student answers in the OSU data set are 11.4 tokens long on average, those in the KU data set have an average length of 5.4 tokens.

Table 1 reports the agreement observed for the meaning assessment. For the detailed and binary assessment, we report the percentage agreement as well as Cohen's Kappa (Cohen, 1960) for the two annotations of the student answers.

	# Student Answers	%		κ	
		binary	detailed	binary	detailed
KU:	5257	88.5%	86.6%	0.712	0.771
OSU:	4826	85.7%	70.6%	0.572	0.473

Table 1: Inter-annotator agreement for meaning assessment

We also grouped the data records by questions (which have unique IDs) and computed the macro average of the respective percentage agreement and Kappa values. The results are shown in Table 2. The purpose of the macro average is to even out the impact of questions with a high number of student answers, which naturally have a high impact on regular, micro-averaged results.

	# Student Answers	# Questions	%		κ	
			binary	detailed	binary	detailed
KU:	5257	202	90.1%	85.2%	0.752	0.745
OSU:	4826	142	86.2%	70.7%	0.583	0.476

Table 2: Macro-averaged inter-annotator agreement for meaning assessment

The macro average makes questions with only few answers equally important as those with many answers, hence reducing the effect of the particularities of few questions with many answers and their meaning assessment annotation.

Computing averaged Kappas is not as straightforward as computing the averages of single Kappa values. Consider the formula proposed by Cohen (1960):

$$\kappa = \frac{p_0 - p_c}{1 - p_c}$$

where p_0 is the probability of the two annotators agreeing based on the observations, also called percentage agreement. p_c refers to the probability of the annotators agreeing by chance, based on the distribution of the labels they used in the input data. For building a macro average, p_c must be computed for the entire data set, whereas p_0 is substituted by the averaged accuracy values, resulting in this slightly modified formula

$$\kappa = \frac{\overline{p_0} - p_c}{1 - p_c}$$

where $\overline{p_0}$ is the average of accuracies grouped by questions. Note that this is different from other approaches to pooled Kappa statistics as, e.g., De Vries et al. (2008), who assume that measures based on different labels and different data sets are to be summarized, not subsets as in our case. Chklovski and Mihalcea (2003) compute averaged Kappa values without using a global p_c, which we consider not appropriate for our setting, since the grouping by questions is made in data analysis afterwards, it is not one that the annotators are aware of as they work on exercise sheet by exercise sheet.

The percentage agreement figures in Table 1 show that both the KU and the OSU team performed well, with all figures above 85% except for the detailed agreement in the OSU annotators. The macro-average figures differ only slightly, hence it is unlikely that single questions with many student answers result in a biasing effect to the other summary statistics.

While the Kappa values above 0.71 in Table 1 exhibit fairly good agreement for both detailed and binary assessment for the KU team, the values for the OSU team range between 0.47 and 0.58, thus indicating that improvements need to be made and further investigation is necessary. As before, the macro-averages differ only slightly.

In order to investigate the problematic Kappa values, we took a closer look at the distribution of labels used by the two annotators. These corresponding figures are shown in Tables 3 and 4.

KU		
	appropriate	*inappropriate*
A1:	76.7% (4032)	23.3% (1225)
A2:	68.8% (3619)	31.2% (1638)

OSU		
	appropriate	*inappropriate*
A1:	73.1% (3528)	26.9% (1298)
A2:	86.3% (4166)	13.7% (660)

Table 3: Distribution of categories used in *binary* assessment by annotators A1 and A2

KU					
	correct answer	*extra concept*	*missing concept*	*blend*	*non-answer*
A1:	59.8% (3145)	2.2% (116)	14.1% (742)	23.0% (1209)	0.7% (45)
A2:	57.6% (3028)	3.6% (189)	14.1% (739)	23.1% (1213)	1.7% (88)

OSU					
	correct answer	*extra concept*	*missing concept*	*blend*	*non-answer*
A1:	57.8% (2791)	5.8% (281)	25.4% (1224)	10.7% (516)	0.3% (14)
A2:	63.8% (3078)	4.2% (203)	26.5% (1280)	2.7% (132)	2.8% (133)

Table 4: Distribution of categories used in *detailed* assessment by annotators A1 and A2

As previously discussed for the Kappa formulas, the percentage agreement is equivalent to the agreement probability p_0. In the case of binary assessment, the two data sets differ only slightly in p_0. Larger values of the chance probability p_c reduce the Kappa values. For binary assessment, we

computed $p_c = 0.60$ for the KU data set and $p_c = 0.67$ for the OSU data set. As Di Eugenio and Glass (2004) point out, this is a problem of skewed categories in the data, an issue which they call the *prevalence problem*. As is apparent from Table 3, the OSU data for binary assessment are even more skewed than those from KU. In the case of detailed assessment, however, there is less of a difference between $p_c = 0.42$ for KU and $p_c = 0.44$ for OSU. Yet, the lower p_0 values in the OSU data (percentage agreement of 70.6%) make the Kappa value drop.

This indicates that the annotators differ in their understandings of how the detailed categories should be assigned. It is not clear-cut if this is due to annotator bias, quality, or to the OSU data being more difficult to work with, e.g., given the significantly longer average length of the OSU learner answers. The data sets collected at the two universities may also differ in terms of the proficiency of the learners from which the data was collected; no standardized proficiency test results were collected. To investigate these issues, it would be interesting to extract a balanced subset of the KU data with full agreement in the binary assessment and have it annotated by the OSU team, and vice versa. This would address the problem with skewed distributions, and it would make it possible to reliably compare the two data sets in terms of the difficulty they present to the annotators.

We also computed the overall percentage agreement for the selection of target answers for those questions with more than one target answer. For the OSU data, of the 142 overall questions (which 4826 overall learner answers), 73 questions had more than one target answer (with 2198 learner answers). For those 2198 learner answers, the two annotators agreed on the same target answer in 84.4%. For the KU data, there are 202 overall questions (with 5257 learner answers). 44 questions had more than one target answer (with 843 learner answers). For those 843 learner answers, the two annotators agreed on the same target answer in 92.9%. We found no evidence for a relation between the agreement in binary or detailed classification and the number of available target answers. In other words, the number of target answers of a given question did not seem to be a factor influencing the difficulty of determining a target answer.

To explore the relation between the binary and the detailed assessment, we investigated in which

cases the annotators agreed in detailed classes, but disagree in binary evaluation. Table 5 shows the sources of such binary disagreement. The clear leader in this table is the category *missing concept*.

	correct	*extra concept*	*missing concept*	*blend*	*non-answer*
KU 401 cases 7.63% of answers	0	14 (3.49%)	309 (77.05%)	78 (19.45%)	0
OSU 401 cases 8.31% of answers	0	1 (0.25%)	364 (90.77%)	36 (8.98%)	0

Table 5: Cases with agreement in detailed assessment but binary disagreement

This means that in many cases, missing information in the student answer for one annotator is still good enough for marking the student answer *appropriate*, while for the other annotator, missing information is enough of a reason to judge it as *inappropriate*. The category *blend*, which by definition overlaps with *missing concept* and *extra concept*, comes up second, further emphasizing the same problem. *Extra concept* shows hardly any binary disagreement, and *correct answer* or *non-answer* are not problematic at all.

4.2 Qualitative Analysis

After discussing the overall quantitative results, we now turn to specific sources of disagreement. By way of example, we will point to each problem in turn and propose a solution that specifies how explicit annotation guidelines could be defined to address these issues. In the following examples, **Q** stands for question, **TA** for target answer and **SA** for student answer.

4.2.1 Enumeration Questions

As expected in reading comprehension, our data contains a fair number of questions that ask for an enumeration of items or statements. An example is the question in Figure 4. The question asks for the movies that can be watched in November and the reading text in this case is a movie schedule containing the relevant information. Whereas the target answer gives an exhaustive listing of both relevant movies, the student merely mentions one of them, *Schwarz auf Weiß*. Both annotators rated

Q: Welche Filme kann man im November sehen?
'Which movies can one see in November?'

TA: Im November kann man Schwarz auf Weiß und (500) Days of Summer sehen .
In November can one Schwarz auf Weiß and (500) Days of Summer watch .

SA: Man kann Schwarz auf Weiß sehen
One can Schwarz auf Weiß see

Annotator A1: *inappropriate*, *missing concept*
Annotator A2: *appropriate*, *missing concept*

Figure 4: Enumeration question

this answer with the detailed code *missing concept* but differ in their overall assessment of the answer. While for annotator A1 the provided content is insufficient, annotator A2 is satisfied with this partial answer.

This discrepancy stems from the fact that it is unclear how complete enumerations need to be for an *appropriate* answer. A general solution for this problem is unlikely as it depends heavily on how the question is phrased: Had the question in this case been *Welche Filme kann man alle im November sehen? (Which films can all be seen in November?)*, listing of all movie titles would have been obligatory due to the trigger word *alle (all)*. A refinement of the annotation guidelines providing a general approach to this issue will thus need to take the linguistic material in the question closely into account.

4.2.2 Learner Strategies

Another potential source for misinterpretation and disagreement relates to specific learner strategies. In particular, lifting (the copying of text snippets) appears to be a popular learner strategy given that it avoids making form errors in the reproduction of content. As mentioned in the introduction, especially weaker students make use of this strategy. For the annotator, understanding the strategy is sometimes crucial to answer interpretation, as exemplified in Figure 5.

Here, the question asks for the two cities Heike visits when she is on vacation. An appropriate answer should state that Heike visits Berlin and Eutin, as the target answer does. However, the student apparently states that he himself sometimes visits Eutin, which at first glance fails to

Q: Welche 2 Städte besucht Heike im Urlaub?
'Which 2 cities does Heike visit during vacation?'

TA: Heike besucht Berlin und Eutin.
Heike visits Berlin and Eutin.

SA: Fahre ich manchmal nach Eutin.
Travel I sometimes to Eutin.

Annotator A1: *inappropriate*, *missing concept*
Annotator A2: *appropriate*, *missing concept*

Figure 5: Use of lifting strategy by learner

address the question. More likely than not, this is why annotator A1 decided to mark this answer as *inappropriate*. By looking at the text, however, one can observe that it is written from a first person perspective and includes the sentence *Im Sommerurlaub fahre ich manchmal nach Eutin. (In the summer vacation, I sometimes drive to Eutin.)*, which includes the underlined string copied verbatim by the student. The student thus correctly located the relevant information in the text, but failed to make the transfer from first to third person. For annotator A2, this was still enough to mark the answer as *appropriate* in binary assessment.

It is debatable what should be done in such cases of lifting. On the one hand, the student showed some comprehension of question and text by locating the relevant passage in the reading text. On the other hand, the production task was not carried out correctly as the necessary form and content manipulation did not take place. From the point of view of a teacher who wants to test comprehension, the answer arguably deserves at least partial credit. However, under the perspective of our project, the expressed linguistic meaning of the answer is the primary object of research and thus such answers should likely be labelled as *inappropriate* since learner strategies and skills such as locating relevant information are beyond the scope of linguistic meaning analysis.

Although form errors should generally have no impact on meaning assessment, there are, of course, cases where form errors directly harm comprehension. This raises the interesting question of how to interpret linguistically ill-formed sentences where hypothetical target forms are necessary for interpreting the meaning at all. In these cases, obtaining such target forms should be done in a

consistent manner with the help of explicit guidelines such as those described by Lüdeling et al. (2005).

4.2.3 Multiple Readings

Turning to the propositional content of answers, there also are cases where the answer itself is well-formed but ambiguous in a way that crucially affects meaning assessment. Consider Figure 6, where the student answer has several readings of which not all are appropriate with regard to the question.

Q: Haben alle Zimmer eine Dusche?
'Do all rooms have a shower?'

TA: Nein, nicht alle Zimmer haben eine Dusche.
No, not all rooms have a shower.

SA: Nein, alle Zimmer haben keine Dusche.
No, all rooms have no shower.

Annotator A1: *appropriate*, *extra concept*
Annotator A2: *inappropriate*, *blend*

Figure 6: Negation scope ambiguity in learner answer

The question is a typical yes/no-question that asks whether all rooms have a shower. The rather verbose target answer unambiguously states that this is not the case. The student answer, on the other hand, can be interpreted as "no room has a shower" or as "not all rooms have a shower". More formally, it has the two readings shown in (3).

(3) a. $\neg(\forall x(room(x) \rightarrow has\text{-}shower(x)))$

b. $\forall x(room(x) \rightarrow \neg\, has\text{-}shower(x))$

The crucial difference is the scope of the negation, which in the situation described in the text renders the meaning true in reading (3a) but false in reading (3b). Where annotator A1 based his assessment on the first reading, annotator A2 seems to have understood the answer according to the second reading.

One can see that, although it is an interesting real-life task of clear interest for (computational)

linguistic research, answer assessment is not always consistent enough by itself. To obtain higher inter-annotator agreement, more explicit guidelines would be needed, going one step further away from observation of the assessment carried out by regular teachers to a more consistent but research-tailored annotation. For the present case, the guidelines could state that as long as there is an appropriate reading, an answer should be marked as *appropriate* and detailed assessment should be based on that reading. Having said that, polarity questions such as the one in Figure 6 can simply be answered with "Yes" or "No", so assessment should primarily be based on the polar part of the answer.

5 Meaning Assessment Results

Having discussed inter-annotator agreement in detail in the previous section, we now present an overview of the meaning assessment results based on the data that both annotators agree on.

Table 6 presents the binary assessment results. It is immediately clear that the majority of answers is deemed *appropriate* by the annotators, as can be seen in the 75.73% for the KU data set and the even higher 84.64% in the OSU data set. This suggests that reading comprehension exercises in real-life foreign language teaching are designed so that they can successfully be completed by the majority of the students at a given level.

appropriate	*inappropriate*
KU (4652 answers with binary agreement)	
75.73% (3523)	24.27% (1129)
OSU (4140 answers with binary agreement)	
84.64% (3504)	15.36% (636)

Table 6: Binary assessment results

Turning to the detailed assessment in Table 7, we can see that certain categories are used much more often than others. For example, the *extra concept* category is only used in less than 2% of the cases and the *incompatible* category for non-answers is even more rare at less than 1%.

In addition, we related the binary and detailed assessment types in Table 8 based on data with

correct answer	*extra concept*	*missing concept*	*blend*	*non-answer*
KU (4556 answers with detailed agreement)				
62.99% (2870)	**1.36%** (62)	12.29% (560)	22.89% (1043)	**0.46%** (21)
OSU (3406 answers with detailed agreement)				
73.78% (2513)	**1.61%** (55)	22.28% (759)	1.94% (66)	**0.38%** (13)

Table 7: Detailed assessment results

agreement in both assessment systems. This enables us to observe the distribution of detailed categories given a certain binary decision.

	correct answer	*extra concept*	*missing concept*	*blend*	*non-answer*
	KU (4155 answers with full agreement)				
appropriate	91.11% (2870)	1.52% (48)	7.05% (222)	0.32% (10)	0
inappropriate	0	0	2.89% (29)	**95.02%** (955)	2.09% (21)
	OSU (3005 answers with full agreement)				
appropriate	91.78% (2513)	1.97% (54)	6.03% (165)	0.22% (6)	0
inappropriate	0	0	**86.14%** (230)	8.99% (24)	4.87% (13)

Table 8: Relating both assessments

The table shows that while in the KU data set, 95.05% of the inappropriate answers are labelled with *blend*, a similar portion of inappropriate answers in the OSU data set (86.14%) are labelled with *missing concept*. The discrepancy in detailed categories given negative binary assessment shows that our current detailed categories, though generally applicable in this domain, are often not used in a consistent fashion. Therefore, as also argued for in Meurers, Ziai, Ott, and Bailey (2011), we pursue the development of an alternative detailed meaning assessment scheme which builds on the type of information requested by the question. Since questions differ with respect to the kind of information and the exhaustivity they require (cf., e.g., Beck and Rullmann 1999), it is not particularly surprising that a single set of five meaning assessment categories cannot cover all types of questions. Instead, refinements by question type will have to be done and meaning assessment categories will have to be in line with question types.

6 Avenues for Future Research

In this section, we describe several directions for future research that we plan to follow based on our corpus. They pertain to two main themes, namely the role of the linguistic context in meaning assessment (sections 6.1 and 6.2) and the choice of linguistic representation for meaning comparison (section 6.3).

6.1 The Information Structure of Answers in Reading Comprehension Tasks

Research into information structure as an interface between the sentence-level semantics and the discourse level has long followed the practice of illustrating its notions using explicit question/answer pairs. In particular, **focus** as a formal pragmatic notion is understood as the part of the answer which is *congruent* with the question, i.e., which answers it (cf., e.g., Krifka 2007). It follows that the notion of focus should be of importance to any answer evaluation approach, as it allows a partitioning of the answer into relevant (focused) and irrelevant (unfocused or backgrounded) material with regard to the question under discussion.

Another important notion is **givenness** (Schwarzschild, 1999), which in its simplest form describes the repetition of previously mentioned material, either lexically or referentially. From an answer assessment perspective, given material should therefore be treated specially, since in most cases given material will not contain the information required by the question. However, this does not hold for all cases: Alternative questions such as *Did Paul kiss Mary or did he kiss Jane?* require a choice between several given alternatives, in this case *Mary* and *Jane*.

In other cases, information is new in the answer but unrequested by the question and thus not needed to determine whether the question has been answered. Figure 7 illustrates this with two appropriate answers differing significantly in the new information (shown in italics) they provide. We therefore cannot assume focus to refer to all new answer material but rather to the part of the answer that supplies the information contextually required by the question.

In connection with focused material in answers, we are also interested in the types of information

Q: An was denken viele Menschen, wenn sie von Weißrussland hören?
'What do many people think of when they hear about Belarus?'

TA: Sie denken an *die Tschernobyl-Katastrophe von 1986.*
They think of the Tschernobyl disaster of 1986.

SA: Ausländer denken bei Weißrussland *weniger an Urlaub*, *sondern eher an die Tschernobyl-Katastrophe von 1986. Damals explodierten in der Sowjetunion Teile eines Atomkraftwerks und wurden einige Regionen Weißrusslands von der radioaktiven Strahlung verseucht*.
Foreigners thinking about Belarus think less of vacation, but rather of the Tschernobyl disaster of 1986. Back then, exploded in the Soviet union parts of a nuclear plant and were some areas of Belarus by the radioactivity polluted.

Figure 7: Example with new but unrequested information.

required by questions, as focused material needs to correspond to such types: if the question asks for a person, the focused part in the answer needs to denote a person in order to satisfy the information requirement. Such **answer types** are commonly used in the Question Answering literature (cf., e.g., Li and Roth 2002) in order to guide Question Answering systems and narrow down the space of potential answer candidates.

One promising way of shedding some light on how focus, givenness and answer types interact in authentic data would be to annotate instances of them in the real-life responses of the CREG corpus presented in this paper. However, an annotation effort would only treat the question and the answer as part of the discourse, since the reading text should naturally contain all the information provided in correct answers and hence would render everything given if considered for annotation.

Concerning focus and givenness, an annotation approach is described by Calhoun et al. (2010), who distinguish various types of given information, new information and focused information based on an Alternative Semantics definition of focus (Rooth, 1992). Further work on annotating givenness and newness is described by Dipper, Götze, and Skopeteas (2007) and Riester, Lorenz, and Seemann (2010). Answer types vary heavily with the domain they are used for, so an applicable scheme of types for our purposes will likely have to be constructed specially. Nevertheless, we are planning to build on the prior research into the annotation of information structure and answer types in order to enrich a part of our corpus with such annotation.

6.2 Text Structure and Task Strategies

The task of answering a reading comprehension question can require students to apply different strategies. Various strategies are described in the reading comprehension literature. For example, Day and Park (2005) describe a two-dimensional classification scheme distinguishing the forms of questions in such exercises and the types of comprehension required from the students to answer the questions. In the scheme of Day and Park (2005), question forms, which are also known from the question answering literature (cf., e.g., Harabagiu and Moldovan 2003), are classes such as yes-no questions, wh-questions, true-or-false questions. Comprehension types refer to the task a student must apply in order to answer a reading comprehension question. While question forms depend only on the question, comprehension types also must take into account how the information is presented to the reader in the text.

Meurers, Ziai, Ott, and Kopp (2011) annotated a small part of CREG with a category system based on the one by Day and Park (2005), finding that the corpus only contains the comprehension types *literal*, *reorganization*, and *inference*. This is not surprising, since we focus on the collection of those reading comprehension questions that ask for information encoded in the corresponding reading texts, no world knowledge should be required (see section 2). Hence, the categories *prediction*, *evaluation*, and *personal response* do not apply.

An annotation with a carefully refined system of that kind would allow users of the corpus to investigate a number of research questions. For example, it would be possible to examine if exercises from the advanced levels require students to use advanced techniques that require them to logically combine information in the text (*inference*), whereas beginner levels might require only a smart lifting of material from the surface form of the text (*literal*). Given that exercises with the more challenging comprehension types were given to students of all levels, it would be interesting to see, how they perform on them along with their learning progress, thereby possibly tracking their development from purely manipulating words by lifting to actually conceptualizing meaning and reproducing it in responses.

6.3 Linguistic Annotation and Automatic Meaning Comparison

Linguistic annotation of corpora in general enables researchers to capture generalizations that are not easily accessible directly in the surface forms. It thus supports insightful new perspectives on learner data, such as the analysis of overuse and underuse of constructions by advanced learners (Hirschmann et al., 2012) on the basis of an automatically parsed German learner corpus.

In the same vein, for our project goal of developing an automatic meaning comparison approach on the basis of the corpus presented in this paper, we have developed an architecture which compares learner and target answers to reading comprehension questions at different levels of linguistic abstraction, including surface forms, lemmas, syntactic phrases and dependencies, as well as semantic representations (Meurers, Ziai, Ott, and Bailey, 2011; Meurers, Ziai, Ott, and Kopp, 2011). In terms of the components of such an architecture, Ott and Ziai (2010) discuss the performance of dependency parsing on a subset of the CREG data introduced in this paper. In order to assess the performance, they manually annotated parts of the reading comprehension corpus with dependency structures; the manual dependency annotation of a larger part of CREG is in progress. Based on such a dependency analysis, Hahn and Meurers (2011) show how to derive explicit semantic representations, for which Lexical Resource Semantics (Richter and Sailer, 2004) is used as a flexible framework supporting underspecification.

7 Summary

We motivated the creation of task-based corpora of authentic language data in context, supporting the interpretation of learner language. To that end, we are collecting a large learner corpus of German reading comprehension exercises with a rich structure: linguistic context (questions and reading texts), student data and metadata, teacher targets and meaning assessment. For distributed corpus collection, we developed and presented the WELCOME tool which centrally stores our data and is freely available for research use.

Based on the data collected so far, we presented an inter-annotator agreement study of the

meaning assessment task. Quantitative results show that the KU data set generally features good agreement with all percentage agreement values over 85% and with κ-values above 0.71. While the OSU data set also achieves over 85% percentage agreement on the binary assessment task, it drops to 70.6% for the detailed assessment task, with generally low κ-values between 0.47 and 0.58. For the binary assessment case, the κ suffers from a skewed distribution. In the detailed case, cross-site reannotation of a balanced subset would be needed to gain further insights into the nature and causes of the differences.

We took a brief look at the overall meaning assessment results based on the current data with full agreement. Results show that the majority of answers is deemed appropriate by the annotators in both data sets. A discrepancy was again found between the two data sets in the use of the detailed categories *missing concept* and *blend*, with the KU team preferring the latter and the OSU team the former. This finding provides further motivation for the development of a better detailed category system for meaning assessment, also taking into account different question types.

Finally, we sketched several avenues for further research into meaning assessment based on additional annotation of our corpus. They include the annotation of information structure notions such as focus and givenness, but also the classification of texts with regard to how they encode required information. Last but not least, we plan to add dependency annotation with the goal of building semantic representations suitable for automatic meaning comparison.

We will make the corpus described here freely available under a Creative Commons by-nc-sa license in the hope of stimulating further research into meaning assessment as well as other areas of learner language research and language in context more generally.

Acknowledgments

The research presented in this paper was funded by the German Science Foundation (DFG) as part of project A4 *Comparing Meaning in Context (CoMiC)* in the SFB 833 *The construction of meaning – the dynamics and adaptivity of linguistic structures*. We would like to thank Nina Vyatkina and

her team at the University of Kansas and Kathy Corl and her group at The Ohio State University for their collaboration in collecting and evaluating the learner data.

In terms of this paper, we are grateful to Nina Vyatkina and the anonymous reviewer for their detailed and very helpful comments.

References

Amaral, L. and D. Meurers (2008). From recording linguistic competence to supporting inferences about language acquisition in context: Extending the conceptualization of student models for Intelligent Computer-Assisted Language Learning. *Computer-Assisted Language Learning 21*(4), 323–338. http://purl.org/dm/papers/amaral-meurers-call08.html.

Bachman, L. F. and A. S. Palmer (1996). *Language Testing in Practice: Designing and Developing Useful Language Tests*. Oxford University Press.

Bailey, S. (2008). *Content Assessment in Intelligent Computer-Aided Language Learning: Meaning Error Diagnosis for English as a Second Language*. Ph. D. thesis, The Ohio State University. http://purl.org/net/Bailey-08.pdf.

Bailey, S. and D. Meurers (2008). Diagnosing meaning errors in short answers to reading comprehension questions. In J. Tetreault, J. Burstein, and R. D. Felice (Eds.), *Proceedings of the 3rd Workshop on Innovative Use of NLP for Building Educational Applications (BEA-3) at ACL'08*, Columbus, Ohio, pp. 107–115. http://aclweb.org/anthology/W08-0913.

Beck, S. and H. Rullmann (1999). A flexible approach to exhaustivity in questions. *Natural Language Semantics 7*, 249–298.

Calhoun, S., J. Carletta, J. Brenier, N. Mayo, D. Jurafsky, M. Steedman, and D. Beaver (2010). The NXT-format Switchboard Corpus: A rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. *Language Resources and Evaluation 44*, 387–419.

Chklovski, T. and R. Mihalcea (2003). Exploiting agreement and disagreement of human annotators for word sense disambiguation. In *In Proceedings of RANLP 2003*.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement 20*(1), 37–46.

Day, R. R. and J.-S. Park (2005). Developing reading comprehension questions. *Reading in a Foreign Language 17(1)*, 60–73.

De Vries, H., M. N. Elliot, D. E. Kanouse, and S. S. Teleki (2008). Using pooled Kappa to summarize interrater agreement across many items. *Field Methods 20*(10), 272–282.

Di Eugenio, B. and M. Glass (2004). The Kappa statistic: A second look. *Computational Linguistics 30(1)*, 95–101. http://www.cis.udel.edu/~carberry/CIS-885/Papers/DiEugenio-Kappa-Second-Look.pdf.

Dipper, S., M. Götze, and S. Skopeteas (Eds.) (2007). *Information Structure in Cross-Linguistic Corpora: Annotation Guidelines for Phonology, Morphology, Syntax, Semantics and Information Structure*, Volume 7 of *Interdisciplinary Studies on Information Structure*. Potsdam, Germany: Universitätsverlag Potsdam.

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

Fitzpatrick, E. and M. S. Seegmiller (2004). The Montclair electronic language database project. In U. Connor and T. Upton (Eds.), *Applied Corpus Linguistics: A Multidimensional Perspective*. Amsterdam: Rodopi.

Granger, S. (2008). Learner corpora. In A. Lüdeling and M. Kytö (Eds.), *Corpus linguistics. An international handbook*, pp. 259–275. Berlin, New York: Walter de Gruyter.

Hahn, M. and D. Meurers (2011). On deriving semantic representations from dependencies: A practical approach for evaluating meaning in learner corpora. In *Proceedings of the Intern. Conference on Dependency Linguistics (DEPLING 2011)*, Barcelona. http://purl.org/dm/papers/hahn-meurers-11.html.

Harabagiu, S. and D. Moldovan (2003). Ontologies. In R. Mitkov (Ed.), *The Oxford Handbook of Computational Linguistics*, Chapter 25, pp. 465–482. Oxford University Press.

Hirschmann, H., S. Doolittle, and A. Lüdeling (2007). Syntactic annotation of non-canonical linguistic structures. In *Proceedings of Corpus Linguistics 2007*, Birmingham. http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/neu2/mitarbeiter-innen/anke/pdf/HirschmannDoolittleLuedelingCL2007.pdf.

Hirschmann, H., A. Lüdeling, I. Rehbein, M. Reznicek, and A. Zeldes (2012). Underuse of syntactic categories in falko. a case study on modification. In *20 years of learner corpus research. Looking back, Moving ahead (LCR2011)*. to appear, http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/mitarbeiter-innen-en/marc/LCR2011_proceedings_Hirschmann_Hagen_et_al.pdf.

Krifka, M. (2007). Basic notions of information structure. In C. Fery, G. Fanselow, and M. Krifka (Eds.), *The notions of information structure*, Volume 6 of *Interdisciplinary Studies on Information Structure (ISIS)*. Potsdam: Universitätsverlag Potsdam. http://www.sfb632.uni-potsdam.de/publications/D2/D2_Krifka_a.pdf.

Li, X. and D. Roth (2002, August). Learning question classifiers. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*, Taipei, Taiwan, pp. 1–7. http://aclweb.org/anthology/C02-1150.

Lüdeling, A., M. Walter, E. Kroymann, and P. Adolphs (2005). Multi-level error annotation in learner corpora. In *Proceedings of Corpus Linguistics*, Birmingham. http://www.corpus.bham.ac.uk/PCLC/Falko-CL2006.doc.

Meurers, D., N. Ott, and R. Ziai (2010). Compiling a task-based corpus for the analysis of learner language in context. In *Pre-Proceedings of Linguistic Evidence*, Tübingen, pp. 214–217. http://purl.org/dm/papers/meurers-ott-ziai-10.html.

Meurers, D., R. Ziai, N. Ott, and S. Bailey (2011). Integrating parallel analysis modules to evaluate the meaning of answers to reading comprehension questions. *IJCEELL. Special Issue on Automatic Free-text Evaluation 21*(4), 355–369. http://purl.org/dm/papers/meurers-ziai-ott-bailey-11.html.

Meurers, D., R. Ziai, N. Ott, and J. Kopp (2011, July). Evaluating answers to reading comprehension questions in context: Results for German and the role of information structure. In *Proceedings of the TextInfer 2011 Workshop on Textual Entailment*, Edinburgh, Scotland, UK, pp. 1–9. Association for Computational Linguistics. http://www.aclweb.org/anthology/W11-2401.

Mislevy, R. J. (2006). Cognitive psychology and educational assessment. In R. L. Brennan (Ed.), *Educational Measurement*, ACE/Praeger Series on Higher Education. Praeger Publishers.

Miura, S. (1998). Hiroshima English Learners' Corpus: English learner No. 2 (English I & English II). Department of English Language Education, Hiroshima University. http://purl.org/icall/eigo1.html, http://purl.org/icall/eigo2.html,.

Ott, N. and R. Ziai (2010). Evaluating dependency parsing performance on German learner language. In M. Dickinson, K. Müürisep, and M. Passarotti (Eds.), *Proceedings of the Ninth International Workshop on Treebanks and Linguistic Theories*, Volume 9 of *NEALT Proceeding Series*, pp. 175–186. http://hdl.handle.net/10062/15960.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Richter, F. and M. Sailer (2004). Basic concepts of Lexical Resource Semantics. In A. Beckmann and N. Preining (Eds.), *European Summer School in Logic, Language and Information 2003. Course Material I*, Volume 5 of *Collegium Logicum*, pp. 87–143. Wien: Publication Series of the Kurt Gödel Society.

Riester, A., D. Lorenz, and N. Seemann (2010). A recursive annotation scheme for referential information status. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, Valletta, Malta. http://www.lrec-conf.org/proceedings/lrec2010/summaries/764.html.

Rooth, M. (1992). A theory of focus interpretation. *Natural Language Semantics 1*(1), 75–116.

Schwarzschild, R. (1999). GIVENness, AvoidF and other constraints on the placement of accent. *Natural Language Semantics 7*(2), 141–177.