Short Answer Grading: When Sorting Helps and When it Doesn't

Ulrike Pado and Cornelia Kiefer
HFT Stuttgart
Schellingstr. 24
70176 Stuttgart
ulrike.pado@hft-stuttgart.de

Abstract

Automatic short-answer grading promises improved student feedback at reduced teacher effort both during and after instruction. Automated grading is, however, controversial in high-stakes testing and complex systems can be difficult to set up by non-experts, especially for frequently changing questions. We propose a versatile, domain-independent system that assists *manual* grading by pre-sorting answers according to their similarity to a reference answer. We show near state-of-the-art performance on the task of automatically grading the answers from CREG (Meurers et al., 2011). To evaluate the grader assistance task, we present CSSAG (Computer Science Short Answers in German), a new corpus of German computer science questions answered by natives and highly-proficient non-natives. On this corpus, we demonstrate the positive influence of answer sorting on the slowest-graded, most complex-to-assess questions.

1 Introduction

Recent research on short-answer prompts has focussed mostly on fully automatically predicting student scores (Burrows, Gurevych and Stein (2015)). While research interest has intensified, central problems in practice remain: On a technical note, teachers need to quickly set up reliable automatic grading for frequently changing questions, which is not always feasible for complex systems. An even more basic concern is that the use of an automated system in summative testing (which determines pass or fail or the overall grade for a class) may not be compatible with legal constraints and with student and teacher beliefs about fair grading.

Another issue with short-answer questions themselves is the objectivity of grading – will two different teachers or even the same teacher on two different days award the same number of points to the same answer? Mohler, Bunescu and Mihalcea (2011) present results from the preparation of their test corpus where their judges perfectly agreed on a score 58% of the time, with differences of one point (out of five) in another 23% of cases. This opens a teacher up to justified complaints from students on 19% of questions. Objective, replicable grading therefore is a big concern in teaching, and of course even more so in summative testing. It is also one that can be naturally addressed with the help of automated or semi-automated systems.

We believe that short-answer grading in real-world teaching will not profit most from fully automatic grade prediction. Instead, relatively simple NLP techniques that need little or no domain adaptation to deal with new questions can assist *manual* grading and both improve objectivity and minimize effort.

We present such a grading assistance tool that presents student answers for manual correction ranked by their similarity to the reference answer (or answers). The intuition is that graders will profit from seeing clearly correct and clearly incorrect answers together.

The similarity scores are computed on the lemma level, so that the system is portable to any other language where a lemmatiser exists. Since it only relies on the lexical content of student and reference answer, it is completely independent of a question domain. To further facilitate real-world use, it is packaged as a plugin to the open-source Learning Management System (LMS) Moodle[1] to allow easy use for teachers.

For the purpose of evaluating our system, we in-

[1] www.moodle.org

Ulrike Pado and Cornelia Kiefer 2015. Short answer grading: When sorting helps and when it doesn't. *Proceedings of the 4th workshop on NLP for Computer Assisted Language Learning at NODALIDA 2015*. NEALT Proceedings Series 26 / Linköping Electronic Conference Proceedings 114: 42–50.

troduce Computer Science Short Answers in German (CSSAG), a new data set of nine short-answer questions from the Computer Science domain. Answers were collected from native or near-native speakers and double-annotated (grading conflicts were resolved after annotation by discussion between the annotators). We report our observations about structural differences between the answers to a native-speaker content matter task (as in CSSAG) and a reading comprehension task that primarily tests language skills (as exemplified by the German standard corpus CREG-1032, Meurers, Ziai, Ott and Kopp (2011)).

We evaluate our system in two ways. First, we adapt our ranking task to binary classification and perform classic score prediction (as correct or wrong) for the CREG-1032 and CSSAG data sets. Our shallow tool approximates the state of the art in binary classification for CREG, with a small drop in performance on CSSAG. This shows that the similarity scores carry relevant information for predicting human grades.

Our second evaluation directly addresses our intended task of grader assistance. Time and accuracy data from human graders shows that the ranking of student answers is beneficial especially for questions that are very slow to grade, at no reduction in agreement with gold grades. Further exploration shows that the slow-to-grade questions are worth more points, which indicates that the teacher expects more complex answers. Higher answer complexity entails more difficult grading. Presenting the answers to these questions ranked by similarity to the reference answer results in a simulated speedup of more than 10%.

2 Related Work

The comprehensive overview over the short-answer grading by Burrows et al. (2015) traces the deepening interest in this task over recent years. Burrows et al. identify different eras in short-answer grading represented by clusters of papers that share a common theme. The first short-answer assessment systems worked with the mapping of concepts in student and reference answer. A prominent example is C-Rater (Leacock and Chodorow, 2003), which attempts a rule-based matching of concepts in the student and reference answers. Answers are first normalised on different levels, using, e.g., spell-checking, synonyms and anaphora resolution.

Analogously to trends in general Computational Linguistics, a later important strategy is the use of corpus-based methods that aim to estimate student answer-reference answer similarity from large collections of language data. The first paper from this group describes the Atenea system (Alfonseca and Pérez, 2004; Pérez et al., 2005), which makes use of distributional (vector-space) and surface-based (BLEU) similarity measures derived from large corpora to assess short-answer questions.

Another theme is the use of pattern matching and alignment on different representational levels. As a system for German, an especially relevant example is CoSeC-DE (Hahn and Meurers, 2012). Hahn and Meurers derive underspecified formal semantic representations of question, student and reference answer and use information structure to identify given and new information in the answers. They derive a score based on quality estimates for the alignment of the representations. Their system reaches the highest prediction accuracy for the German standard corpus CREG.

Corpus-based and alignment-based similarity measures are often used as features in the era of machine learning. The machine-learning based paper most relevant for us is CoMiC-DE, the system for German by Meurers et al. (2011). The system uses alignments on various levels of linguistic representation like tokens, chunks, or dependency parses, as well as corpus-based similarity measures to train a memory-based learner. This paper also introduces the CREG corpus, which we further analyse below.

Burrows et al. explicitly define their subject as *automatic* short answer grading, and the vast majority of publications on short answer grading aim for fully automatic grade prediction. We did, however, consciously choose to build an assistance system for *manual* short answer grading.

Two such grader assistance systems have been presented, to our knowledge. Both independently propose the clustering of answers; grading then proceeds per cluster instead of per answer to reduce manual effort. Basu, Jacobs and Vanderwende (2013) use machine learning to train a model of similarity between student answers using vector-based similarity and lexical match features. These similarity scores are then used to hierarchically cluster the answers, allowing teachers to grade multiple answers at the same time and provide detailed feedback on classes of (pos-

sibly erroneous) answers. Basu et al. show that their system reaches 92.9% accuracy in automated binary classification on their 10-question English content-assessment data set. They also find a drastic reduction in the number of actions a grader has to take in order to grade all student answers: 40-50% of simulated actions can be saved to reach the same grading result as answer-by-answer grading.

In a follow-up paper, Brooks, Basu, Jacobs and Vanderwende (2014) present a user study for the system. Overall, teachers were able to assign a grade to every answer three times as quickly with the system, while their agreement with the gold score did not suffer.

Horbach, Palmer and Wolska (2014) cluster student short answers flatly using surface features (word and character n-grams, presence of predefined core keywords). They make the explicit assumption that a small number of incorrectly graded items is acceptable as long as the teacher's workload is greatly reduced. They evaluate on German learner listening comprehension material: Using their system, a simulated teacher can reach 85% agreement with the gold score by labelling only 40% of responses.

3 The Grader Assistance System

Our system relies on determining the similarity of student and reference answer and then sorting the student answers according to this similarity. In contrast to Horbach et al. (2014) and Basu et al. (2013), we do not cluster student answers, because teachers need to see every single answer in order to make the tool acceptable for use in summative assessment.

The similarity score is computed on filtered lemmas from the student and reference answer. Further, we demote words from the question (Mohler et al. (2011)) to only retain content word lemmas that are relevant to the new content in the student or reference answer. This is a shallow approximation of content rather than surface form. Note, however, that so far, we do not include synonyms nor handle paraphrases. At this point, our goal was to evaluate a very simple, versatile system which does not need domain adaptation.

Table 1 shows the processing steps for an example question. The analysis system uses the DKPro Core (de Castilho and Gurevych, 2014) and DKPro Similarity (Bär, Zesch and Gurevych (2013)) libraries. We compute lemmas using the Stanford lemmatiser component in DKPro Core (Manning, Surdeanu, Bauer, Finkel, Bethard and McClosky (2014)) and exclude stop words using Porter's German stop word list[2]. We then exclude all lemmas from the student and reference answer that already appear in the question. The similarity between student and reference answer is calculated using the DKPro Similarity implementation of Greedy String Tiling (as proposed by Wise (1996)). This text similarity measure aims to find (the longest possible) matching substrings, regardless of position in the original text, and ranges between 0 (no match) and 1 (perfect agreement).

If more than one reference answer is provided, the similarity of the student answer to all variants of reference answers will be calculated and the highest score will be used.

The system is implemented as a plugin to the LMS Moodle[3] and available under the GPL. The implementation can easily be ported to other LMS, as well.

4 CSSAG (Computer Science Short Answers in German)

We collected a data set of nine short-answer questions and answers collected over the course of a one-semester Introduction to Programming in Java class aimed at first-year undergraduate students. The questions test students' knowledge of basic object-oriented programming concepts. In week 5, for example, students had to explain the relationship between classes and objects (German question: "Erklären Sie den Zusammenhang zwischen Klassen und Objekten."). Students are native speakers of German or have sufficient German skills to pursue higher education exclusively in German.

There are a total of 491 answers, with an average of 55 answers per questions (min 33, max 83) and at least one reference answer meant for human graders. Answers are one to three sentences long.

Each question was graded (out of one or two points in increments of 0.5 points) by two experienced graders who are domain experts. Cases of disagreement were adjudicated by discussion between the graders; when necessary, the reference answer was disambiguated or extended. No an-

[2] http://snowball.tartarus.org/algorithms/german/stop.txt

[3] https://github.com/HftAssistedGrading/moodle-plugin-assisted-grading

Frage: Erklären Sie den Zusammenhang zwischen Klassen und Objekten
Question: Explain the relationship between classes and objects.

	Reference Answer	Student Answer
Original	Eine Klasse ist der Bauplan für ein Objekt. Ein Objekt ist eine konkrete Instanz einer Klasse. *A class is the blueprint for an object. An object is a concrete instance of a class.*	Eine Klasse ist ein Bauplan für ein Objekt. Die Klasse definiert den Typ des Objektes. Ein Objekt ist eine Ausprägung *A class is a blueprint for the object. The class defines the type of the object. An object is a realisation.*
Lemmas, no stopwords	klasse bauplan objekt objekt konkret instanz klasse	klasse bauplan objekt klasse definieren typ objekt objekt ausprägung
Question demotion	bauplan konkret instanz	bauplan definieren typ ausprägung

Table 1: Processing steps in the Grader Assistance system.

swers were excluded from the corpus.

We intend to make the data set publicly available for research.

5 Experiment 1: Binary Classification

Our evaluation is two-fold: Our first experiment establishes that the similarity between student and reference answers does indeed predict human-assigned grades. We then go on to test the influence of ranking the student answers on grading speed and agreement with the gold grade.

In Experiment 1, we classify student short answers as correct or incorrect given their similarity to the reference answer. This is the classical automatic short answer grading task given a two-level scoring regime. We compare our results against Hahn and Meurers (2012) who report the best results to date on CREG, the German short-answer corpus. Their system runs a deep semantic analysis to derive underspecified formal semantic representations of the question, student and reference answer and determine information structural focus.

5.1 Data

In addition to CSSAG, we use the CREG-1032 corpus as described in Meurers et al. (2011). It contains German learner answers to reading comprehension questions.

5.2 Method

We use the similarity scores to classify answers as correct or wrong by determining a similarity threshold. Scores above the threshold are taken to indicate a correct answer (due to its large similarity to the reference answer), scores below the threshold are counted as incorrect answers.

System	CREG	CSSAG
Frequency Baseline	50.0	64.6 (strict) 53.4 (generous)
Grader Assistance	83.7	78.0 (strict) 80.0 (generous)
Meurers et al. ('11)	84.7	–
Hahn&Meurers ('12)	86.3	–

Table 2: Exp. 1: Results for binary classification by the Grader Assistance system on the CREG-1032 and CSSAG data sets.

The threshold was set a priori at 0.49 as the mid-point of the similarity scale. The value was checked for plausibility on a held-out question from the CSSAG data set (question ID w4). The threshold was, however, not optimised for either corpus, so further improvements may be possible when the threshold is adapted. Empirically setting the threshold poses the interesting problem of sampling a representative development set, since the set should not overlap with the test data and there is considerable variation between the different questions.

The CREG data set can be evaluated right away given the threshold, as answers are either fully correct or incorrect. On the CSSAG data set, partial credit was awarded. We therefore report two scoring methods: *strict* scoring counts only answers with full points as correct, *generous* scoring counts answers with full or partial points as correct.

5.3 Results and Discussion

Table 2 reports the results. We compare the systems against the frequency baseline for each data set (i.e., the prediction accuracy for always predicting the most frequent class). The CREG data are constructed to contain exactly half correct and half incorrect answers, so the frequency baseline on this data set is 50%. On the CSSAG data, the bias of the scoring methods is clearly visible: The strict method only counts those answers as correct that were assigned full points. About two thirds of the student answers are consequently classed as incorrect, and the frequency baseline (when predicting "incorrect") is much higher than for the generous scoring method, where answers with partial points also count as correct. For generous scoring, the frequency baseline is close to 50%.

Our grading assistance system reaches roughly 84% accuracy on the CREG-1032 data set. This comes close to the best result to date, 86.3% reported for the deep Hahn and Meurers (2012) CoSeC-DE system. Our shallow analysis is thus able to roughly approximate the state-of-the-art. Apparently, the corpus contains only a small portion of answers that are graded incorrectly by the shallow method and need to be deeply analysed for accurate scoring. We discuss this observation further in Section 5.4.

On the CSSAG data set, the system accuracy reaches 78% for strict and 80% for generous interpretation of partial points. While these numbers are noticeably lower than on the CREG data set, the system clearly outperforms the frequency baselines. It gains noticeably more over the generous baseline than over the strict baseline: It appears to be easier for our simple string similarity strategy to distinguish between wrong and (partially) correct answers than to tell apart partially correct and fully correct answers. In any case, the results imply a meaningful relation between similarity to the reference answer and human-assigned grade.

5.4 Corpus Comparison

Further analysis of the test corpora revealed interesting differences in their characteristics. We find that the correct answers in CREG are generally very similar to the reference answer, markedly more so than for the CSSAG data.

To estimate the variance within the answers, we report the average similarity score between student and reference answers as computed by our system in Table 3. For CREG, answers have an average similarity score of 0.39 to the reference answer. This number even goes up to 0.65 for just the correct answers. With the CSSAG corpus, the average score over all answers is much lower at 0.27 (or 0.54 for the answers with full gold scores).

Corpus	All Questions	Correct Questions
CREG	0.39	0.65
CSSAG	0.27	0.54

Table 3: Corpus comparison: Average similarity of student answers to reference answer in CREG and CSSAG corpora. CSSAG correct answers by strict interpretation of points assigned.

The high similarity of correct student answers to the reference answer explains the success of our shallow method in classifying CREG answers: Simple string matching to the reference easily reveals the correct answers.

In general, the higher CREG similarity scores indicate much less variance among the answers in CREG than in CSSAG. This empirical finding is at odds with the usual theoretical assumptions about short-answer questions: Limited answer variance is a hallmark of closed question types like fill-in-the-blank, while short answer questions are seen as an open question type with generally high answer variation. Our results imply that within a theoretically open question type, there is a range of actual answer variation. To our knowledge, this observation is new in the literature, although it clearly has repercussions for automatic grading or grading assistance, with more open questions being more difficult to treat. Evaluation results should therefore be interpreted in the context of answer variation in the test data: The results that can be expected from deep and shallow models respectively depend on the amount of variation in the answers relative to the reference answer, with little variation favouring shallow models.

One contributing factor to the closedness of CREG questions is that the corpus contains only answers that were graded consistently by all annotators. This means that the classification as correct or incorrect is very certain, but the distinction is artificially made more clear-cut than it really is. CSSAG in contrast contains all available student answers, with grader inconsistencies addressed by grader discussion after the initial annotation.

Apart from design decisions, there are also linguistic and psycholinguistic reasons for more answer variation in CSSAG: There is a difference both in tasks and student population. In the reading comprehension task reflected in the CREG data, students have all recently read the same text and are presumably primed by its lexical and syntactic features (Meyer and Schvaneveldt, 1971; Bock, 1986). This means they are more likely to use the same words and structures in their answers (Pickering and Garrod, 2004), even if explicit answer lifting (copying from the text) is not considered. In addition, learners may lack the vocabulary and language skills to paraphrase freely. In contrast, the CSSAG questions assess mastery of content taught several days previously, and the students are mostly native speakers, with the non-natives skilled enough to pursue higher education exclusively in German. This student pool produces a wider range of paraphrases of the correct answer.

In sum, the high similarity of the correct CREG answers allows a rough content matching algorithm such as ours to reach the performance of linguistically more complex systems. An interesting question for further research is to evaluate the performance of the more complex systems on the CSSAG data set. With more varied answer phrasing, the complex strategies may show more pronounced gains.

6 Experiment 2: Agreement and Speed in Grading

Our second evaluation tests the influence of similarity ranking on grading accuracy and speed. This is the task for which we designed the system.

6.1 Method

We presented a group of twelve graders with all questions, reference answers and student answers from CSSAG. Four graders were highly experienced, the other eight were novice graders, but all were knowledgeable in the domain.

The answers were either ordered randomly or sorted according to their similarity to the reference answer. Each grader saw roughly half the questions in sorted and half in random order. This means that each question was annotated by six graders in each of the two conditions. Graders were not informed that some of the answer sets had been sorted, and sorted and random answer sets were in chance order in the work packages. Graders were timed for each question. We had to discard the times for one grader because they were registered incorrectly. We then computed grader agreement with the gold grade and average grading time per answer.

6.2 Results and Discussion

Average agreement of the points assigned by the graders to the gold grades (gold agreement) was comparable between the two groups of graders, although, not surprisingly, the expert graders did somewhat better at 75.7% agreement, compared to 73.4% for the novice graders[4]. The novices took roughly 1.4 times longer for grading than the experts (14.7 vs. 10.5 seconds per answer).

Comparing the random and sorted conditions averaged across all graders yields an interesting picture: Sorting has only a small positive effect on grader agreement with the gold grade at 73.8% agreement in the random condition and 74.6% in the sorted condition. Average grading time is identical (13.6 seconds random, 13.6 seconds sorted).

The grading agreements for the same question across conditions are highly correlated (Spearman's $\rho = 0.92, p < 0.01$) with similar means, implying that there is no influence of our sorting scheme on grader agreement.

In the random condition, there is a significant negative correlation between grading time and agreement ($\rho = -0.795, p < 0.02$), so that answers that are graded more accurately are also quicker to grade. However, this correlation does not hold in the sorted condition. Also, the grading time for the two conditions is not significantly correlated ($\rho = 0.588, p = 0.08$) despite the equal averages. This shows that the sorting does have an effect on grading time, even though the effect appears to be zero-sum, since it does not show in the condition average.

In the sorted condition, we find a significant correlation between grading time and the average similarity score for a question ($\rho = 0.798, p < 0.02$) instead of the correlation between time and agreement in the random condition. Given sorting, a question will be graded faster if the average answer similarity to the reference is low. This is the case for example if there are a great number of fragment answers that are easy to score (as incor-

[4] "Novice" only refers to grading experience; all graders were knowledgeable in the question domain

Question ID	Random	Sorted	Δ
w7	8	12	4
w10	10	13	3
w4	11	9	-2
w6	12	12	0
pvl2	12.4	14	1.6
pvl3	13.2	11	-2.2
w5	13.5	21.4	7
Average	13.6	13.6	0
w9	**20**	**14**	**-6**
w11	**22**	**16**	**-5**

Table 4: Exp. 2: Average grading times in seconds per condition. Slowest questions gain most from sorting.

Question ID	Δ s	Points
pvl2	1	1.6
w4	1	-2
w5	1	7
w6	1	0
w7	1	4
w10	1	3
pvl3	**2**	**-2.2**
w9	**2**	**-6**
w11	**2**	**-5**

Table 5: Exp. 2: Maximum points per question and time difference between random and sorted conditions. Two-point questions (in bold) show speedup (negative difference).

rect). Scoring all of these answers together seems to free up time for grading the answers with relevant content.

Table 4 shows which questions profit mostly from sorted presentation: We list the average grading times for each question in the random and sorted conditions and the time difference between conditions. The average grading time over all questions is given, as well. The lines in the table are sorted by grading duration in the random condition. The table suggests that the answers that are slower than average to grade gain most by a sorted presentation (by about five seconds per answer). Answers that are faster than average to grade or take average time do not profit from sorting.

This implies that optimally, we should present questions that will be slow to grade in sorted order, and questions that will take average time or less in random order. This raises the question of how to identify the slow-to-grade questions beforehand. Further scrutiny of the questions reveals that speedup by sorting is achieved mainly for those questions where students can earn two points (rather than one, as for the majority of questions). Table 5 shows the questions with their maximum number of points to be earned and the time difference achieved by sorting (a negative difference is a speedup).

Choosing the presentation mode according to the points students can gain for each question has the advantage of relying only on information that is available for every question out of the logic of the task, so no further manual or automatic processing of the questions is required.

Two-point questions differ from one-point questions in the cognitive load on the grader: When creating the question, the teacher already expected complex answers with several facets that are each worth partial points. The grader needs to keep track of all expected and actually given aspects of the answer in order to arrive at the final score. In this cognitively demanding situation, sorting yields speed gains of 15-30% per question.

If the questions had been presented optimally to our graders (answers to one-point questions in random order, answers to two-point questions in sorted order), the average overall grading time per answer would be 12 seconds (based on the experimental by-question averages from the sorted and random conditions). This is a 12% gain, equivalent to 13 minutes saved when grading the total of 491 answers. Agreement with gold grades would be virtually unaffected at an average 73.5% across all questions (as opposed to an average 73.8% in the random condition). Further, only presenting the answers to some questions in sorted order should also help to avoid graders' possible overreliance on the similarity score for grading once they become aware of the sorted presentation. Future work will test the efficacy of the hybrid presentation mode in practice.

7 Conclusions and Future Work

In this paper, we have presented a system to assist manual grading of short-answer questions by ranking student answers in order of their similarity to the reference answer. The system is designed to be domain-independent and easy to use for teachers without computational linguistics expertise. Beside portability and usability, our main

goal was to speed up grading and improve objectivity. Our approach ensures that the teacher still sees every student answer, which is an important prerequisite for use of the system in summative testing.

To evaluate our system, we have introduced a new data set, Computer Science Short Answers in German (CSSAG). The data demonstrably differs from the standard German short-answer corpus CREG (Meurers et al., 2011) in several respects: The questions assess content mastery rather than language skills and were collected from native German speakers. We find that the difference in task and student population make CSSAG answers more variable than CREG answers. Further work will investigate equivalent English corpora.

In our evaluation of the automatic grading task, our shallow tool approximates the state of the art in binary classification for CREG, with a small drop in performance on CSSAG. This shows that the similarity scores carry relevant information for predicting human grades. We also hypothesise that the lower answer variation in CREG makes it easier to automatically grade with a shallow system such as ours. Future work should aim to determine whether more complex systems show more performance gains on CSSAG.

Time data from human graders indicates that the ranking of student answers is beneficial especially for questions that are very slow to grade. These are questions with a maximum grade of more than one point, which reflects their greater complexity and, in consequence, the greater cognitive load on the grader. Optimal answer presentation guided by the maximum number of points that can be earned for each question speeds up grading by 1.6 seconds per answer on average, at undiminished agreement with gold. This simulated result needs to be evaluated experimentally in the future.

References

Enrique Alfonseca and Diana Pérez. 2004. Automatic assessment of open ended questions with a BLEU-inspired algorithm and shallow NLP. In *Advances in Natural Language Processing*, volume 3230 of *Lecture Notes in Computer Science*. Springer.

Daniel Bär, Torsten Zesch, and Iryna Gurevych. 2013. DKPro Similarity: An open source framework for text similarity. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 121–126.

Sumit Basu, Chuck Jacobs, and Lucy Vanderwende. 2013. Powergrading: A clustering approach to amplify human effort for short answer grading. *Transactions of the Association for Computational Linguistics*, 1:391–402.

Kathryn Bock. 1986. Syntactic persistence in language production. *Cognitive Psychology*, 18:355–387.

Michael Brooks, Sumit Basu, Charles Jacobs, and Lucy Vanderwende. 2014. Divide and correct: Using clusters to grade short answers at scale. In *Learning @ Scale*.

Steven Burrows, Iryna Gurevych, and Benno Stein. 2015. The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education*, 25:60–117.

Richard Eckart de Castilho and Iryna Gurevych. 2014. A broad-coverage collection of portable NLP components for building shareable analysis pipelines. In *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT (OIAF4HLT) at COLING 2014*, pages 1–11.

Michael Hahn and Detmar Meurers. 2012. Evaluating the meaning of answers to reading comprehension questions: A semantics-based approach. In *Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications (BEA7)*.

Andrea Horbach, Alexis Palmer, and Magdalena Wolska. 2014. Finding a tradeoff between accuracy and rater's workload in grading clustered short answers. In *Proceedings of the 9th LREC*, pages 588–595.

Claudia Leacock and M. Chodorow. 2003. C-rater: Automated scoring of short-answer questions. *Computers and the Humanities*, 37(4):389–405.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

Detmar Meurers, Ramon Ziai, Niels Ott, and Janina Kopp. 2011. 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*, pages 1–9.

David E. Meyer and Roger W. Schvaneveldt. 1971. Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, 90(2):227–234.

Michael Mohler, Razvan Bunescu, and Rada Mihalcea. 2011. Learning to grade short answer questions using semantic similarity measures and dependency

graph alignments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 752–762. ACL.

Diana Pérez, Enrique Alfonseca, Pilar Rodríguez, Alfio Gliozzo, Carlo Strappavara, and Bernardo Magnini. 2005. About the effects of combining latent semantic analysis with natural language processing techniques for free-text assessment. *Revista Signos: Estudios de Linguïstica*, 38(59):325–343.

Martin Pickering and Simon Garrod. 2004. The interactive-alignment model: Developments and refinements. *Behavioral and Brain Sciences*, 27:212–225.

Michael J. Wise. 1996. YAP3: Improved detection of similarities in computer program and other texts. In *SIGCSEB: SIGCSE Bulletin (ACM Special Interest Group on Computer Science Education)*, pages 130–134. ACM Press.