Can Natural Language Processing Become Natural Language Coaching?

Marti A. Hearst
UC Berkeley
Berkeley, CA 94720
hearst@berkeley.edu

Abstract

How we teach and learn is undergoing a revolution, due to changes in technology and connectivity. Education may be one of the best application areas for advanced NLP techniques, and NLP researchers have much to contribute to this problem, especially in the areas of learning to write, mastery learning, and peer learning. In this paper I consider what happens when we convert natural language processors into natural language coaches.

1 Why Should You Care, NLP Researcher?

There is a revolution in learning underway. Students are taking Massive Open Online Courses as well as online tutorials and paid online courses. Technology and connectivity makes it possible for students to learn from anywhere in the world, at any time, to fit their schedules. And in today's knowledge-based economy, going to school only in one's early years is no longer enough; in future most people are going to need continuous, lifelong education.

Students are changing too — they expect to interact with information and technology. Fortunately, pedagogical research shows significant benefits of active learning over passive methods. The modern view of teaching means students work actively in class, talk with peers, and are coached more than graded by their instructors.

In this new world of education, there is a great need for NLP research to step in and help. I hope in this paper to excite colleagues about the possibilities and suggest a few new ways of looking at them. I do not attempt to cover the field of language and learning comprehensively, nor do I claim there is no work in the field. In fact there is quite a bit, such as a recent special issue on language learning resources (Sharoff et al., 2014), the long running ACL workshops on Building Educational Applications using NLP (Tetreault et al., 2015), and a recent shared task competition on grammatical error detection for second language learners (Ng et al., 2014). But I hope I am casting a few interesting thoughts in this direction for those colleagues who are not focused on this particular topic.

2 How Awkward

Perhaps the least useful feedback that an instructor writes next to a block of prose on a learner's essay is `'awkward'`. We know what this means: something about this text does not read fluently. But this is not helpful feedback; if the student knew how to make the wording flow, he or she would have written it fluently in the first place! Useful feedback is *actionable*: it provides steps to take to make improvements.

A challenge for the field of NLP is how to build writing *tutors* or *coaches* – as opposed to *graders* or *scorers*. There is a vast difference between a tool that performs an assessment of writing and one that coaches students to help them as they are attempting to write.

Current practice uses the output of scorers to give students a target to aim for: revise your essay to get a higher score. An alternative is to design a system that watches alongside a learner as they write an essay, and coaches their work at all levels of construction – phrase level, clause level, sentence level, discourse level, paragraph level, and essay level.

Grammar checking technology has been excellent for years now (Heidorn, 2000), but instead of just showing the right answer as grammar checkers do, a grammar coach should give hints and scaffolding the way a tutor would – not giving the answer explicitly, but showing the path and letting the learner fill in the missing information. When the learner makes incorrect choices, the parser

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, pages 1245–1252,
Beijing, China, July 26-31, 2015. ©2015 Association for Computational Linguistics

can teach principles and lessons for the conceptual stage that the learner is currently at. Different grammars could be developed for learners at different competency levels, as well as for different first-second language pairings in the case of second language learning.

This suggests a different approach for building a parser than what is the current standard. I am not claiming that this has not been suggested in the past; for instance Schwind (1988) designed a parser to explain errors to learners. However, because of the renewed interest in technology for teaching, this may be a pivotal time to reconsider how we develop parsing technology: perhaps we should think fundamentally about parsers as coaches rather than parsers as critics.

This inversion can apply to other aspects of NLP technology as well. For instance, Dale and Kilgarriff (2011) have held a series of workshop to produce algorithms to identify errors introduced into texts by non-native writers in the warmly named "Helping Our Own" shared task (Dale et al., 2012). Using the technology developed for tasks like these, the challenge is to go beyond recognizing and correcting the errors to helping the writer understand why the choices they are making are not correct. Another option is to target practice questions tailored for learners based on errors in a fun manner (as described below).

Of course, for decades, the field of Intelligent Tutoring Systems (ITS) (VanLehn, 2011) has developed technology for this purpose, so what is new about what I am suggesting? First, we know as NLP researchers that language analysis requires specific technology beyond standard algorithms, and so advances in Intelligent Tutoring Systems on language problems most likely requires collaboration with experts in NLP. And, apparently such collaborations have not been as robust as they might be (Borin, 2002; Meurers, 2012). So there is an opportunity for new advances at the intersection of these two fields.

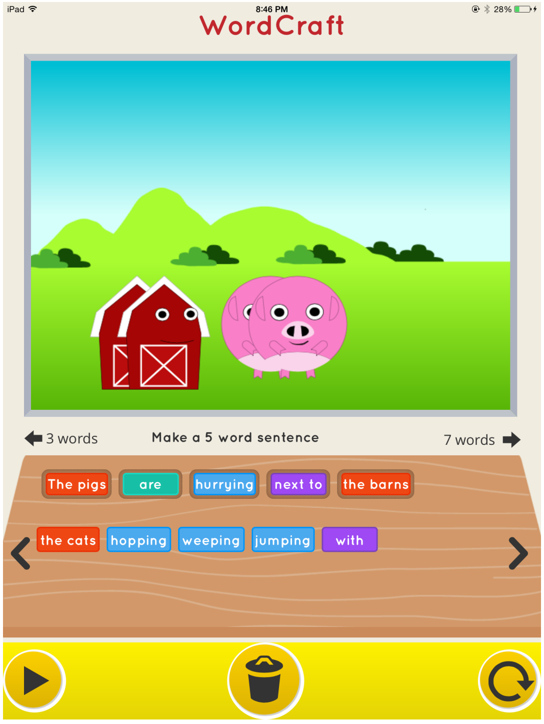

Figure 1: Wordcraft user interface showing a farm scene with four characters, a fully formed sentence, the word tray with candidate additional words colored by part of speech, and tool bar. When the child completes a sentence correctly, the corresponding action is animated.

And second, the newly expanded interest in online learning and technology makes possible the access of information about student writing behavior on a large scale that was not possible in the past. Imagine thousands of students in cascaded waves, tasked with writing essays on the same topic, and receiving real-time suggestions from different algorithms. The first wave of student responses to the feedback would be used to improve the algorithms and these results would be fed into the next wave of student work, and so on. Students and instructors could be encouraged to give feedback via the user interface. Very rapid cycles of iteration should lead to accelerated improvements in understanding of how the interfaces and the algorithms could be improved. A revolution in understanding of how to coach student writing could result!

Algorithms could be designed to give feedback for partially completed work: partially written sentences in the case of a parser; partially completed paragraphs in the case of a discourse writing tool, and so on, rather than only assessing completed work after the fact.

3 Karaoke Anyone?

Beyond learning to write, new technology is changing other aspects of language learning in ways that should excite NLP researchers. In order to write well, a student must have a good vocabulary and must know syntax. Learning words and syntax requires exposure to language in many contexts, both spoken and written, for a student's primary language was well as for learning a second language.

Although computerized vocabulary tools have been around for quite some time, the rise of mobile, connected applications, the serious games movement, and the idea of "microtasks" which are done during interstices of time while out and about during the day, opens the door to new ways to expose students to repetitive learning tasks for acquiring language (Edge et al., 2011). Some of the most innovative approaches for teaching language combine mobile apps with multimedia information.

For example, the Tip Tap Tones project (Edge et al., 2012) attempts to help learners reduce the the challenge of mastering a foreign phonetic system by microtasking with minute-long episodes of mobile gaming. This work focuses in particular on helping learners acquire the tonal sound system of Mandarin Chinese and combines gesture swipes with audio on a smartphone.

The ToneWars app (Head et al., 2014) takes this idea one step farther by linking second language learners with native speakers in real time to play a Tretis-like game against one another to better learn Chinese pronunciation. The second language learner feels especially motivated when they are able to beat the native speaker, and the native speaker contributes their expert tone recordings to the database, fine-tunes their understanding of their own language, and enjoys the benefits of tutoring others in a fun context.

Going beyond phonemes, the DuoLingo second-language learning application (von Ahn, 2013) teaches syntax as well as vocabulary through a game-based interface. For instance, one of Duolingo's games consists of a display of a sentence in one language, and a jumbled list of words in the opposing language presented as cards to be dragged and dropped onto a tray in the correct order to form a sentence. In some cases the user must select between two confounding choices, such as the articles "le" or "la" to modify French nouns.

Our work on a game for children called WordCraft takes this idea one step further (Anand et al., 2015) (see Figure 1). Children manipulate word cards to build sentences which, when grammatically well formed, come to life in a storybook-like animated world to illustrate their meaning. Preliminary studies of the use of Wordcraft found that children between the ages of 4 and 8 were able to observe how different sentence constructions resulted in different meanings and encouraged children to engage in metalinguistic discourse, especially when playing the game with another child.

A karaoke-style video simulation is used by the Engkoo system to teach English to Chinese speakers (Wang et al., 2012). The interface not only generates audio for the English words, but also shows the lip and facial shapes necessary for forming English words using a 3D simulated model lip-syncing the words in a highly realistic manner. To generate a large number of sample sentences, the text was drawn from bilingual sentence pairs from the web.

These technologies have only become feasible recently because of the combination of multimedia, fast audio and image processing, fast network connectivity, and a connected population. NLP researchers may want to let their imaginations consider the possibilities that arise from this new and potent combination.

4 Closing the Cheese Gap

Salman Kahn, the creator of Kahn Academy, talks about the "Swiss cheese" model of learning in which students learn something only partly before they are forced to move on to the next topic, building knowledge on a foundation filled with holes, like the cheese of the same name (Khan, 2012). This is akin to learning to ride a bicycle without perfecting the balancing part. In standard schooling, students are made to move one from one lesson to the next even if they only got 70, 80, 90% correct on the test. By contrast, *mastery learning* requires a deep understanding, working with knowledge and probing it from every angle, trying out the ideas and applying them to solve real problems.

In many cases, mastery learning also requires practicing with dozens, hundreds, or even thousands of different examples, and getting feedback on those examples. Automation can help with mastery learning by generating personalized practice examples that challenge and interest students. Automatically generated examples also reduce the cost of creating new questions for instructors who are concerned about answer sharing among students from previous runs of a course.

Recently, sophisticated techniques developed in the programming languages field have begun to be applied to automate repetitive and structured tasks in education, including problem generation, solution generation, and feedback generation for computer science and logic topics (Gulwani, 2014).

Closer to the subject at hand is the automated generation of mathematical word problems that are organized around themes of interest to kids, such as “School of Wizardry” (Polozov et al., 2015). The method allows the student to specify personal preferences about the world and characters, and then creates mini “plots” for each word problem by enforcing coherence across the sentences using constraints in a logic programming paradigm combined with hand-crafted discourse tropes (constraints on logical graphs) and a natural language generation step. A sample generated word problem is

> Professor Alice assigns Elliot to make a luck potion. He had to spend 9 hours first reading the recipe in the textbook. He spends several hours brewing 11 portions of it. The potion has to be brewed for 3 hours per portion. How many hours did Elliot spend in total?

Results are close in terms of comprehensibility and solubility to those of a textbook. The project’s ultimate goal is to have the word problems actually tell a coherent story, but that challenge is still an open one. But the programs can generate an infinite number of problems with solutions. Other work by the same research team generated personalized algebraic equation problems in a game environment and showed that students could achieve mastery learning in 90 minutes or less during an organized educational campaign (Liu et al., 2015).

Another way that NLP can help with mastery learning is to aid instructors in the providing of feedback on short answer test questions. There has been significant work in this space (Kukich, 2000; Hirschman et al., 2000). The standard approach builds on the classic successful model of essay scoring which compares the student’s text to model essays using a similarity-based technique such as LSA (Landauer et al., 2000; Mohler and Mihalcea, 2009) or careful authoring of the answer (Leacock and Chodorow, 2003).

Recent techniques pair with learning techniques like Inductive Logic Programming with instructor editing to induce logic rules that describe permissible answers with high accuracy (Willis, 2015). Unfortunately most approaches require quite a large number of students’ answers to be marked up manually by the instructor before the feedback is accurate enough to be reliably used for a given question; a recent study found on the order of 500-800 items per question had to be marked up at minimum in order to obtain acceptable correlations with human scorers (Heilman and Madnani, 2015). This high initial cost makes the development of hundreds of practice questions for a given conceptual unit a daunting task for instructors.

Recent research in Learning at Scale has produced some interesting approaches to improving “feedback at scale.” One approach (Brooks et al., 2014) uses a variation on hierarchical text clustering in tandem with a custom user interface that allows instructors to rapidly view clusters and determine which contain correct answers, incorrect answers, and partially correct answers. This greatly speeds up the markup time and allows instructors to assign explanations to a large group of answers with a click of a button.

An entirely different approach to providing feedback that is becoming heavily used in Massive Open Online Courses is peer feedback, in which students assign grades or give feedback to other students on their work (Hicks et al., 2015). Researchers have studied how to refine the process of peer feedback to train students to produce reviews that come within a grade point of that of instructors, with the aid of carefully designed rubrics (Kulkarni et al., 2013).

However, to ensure accurate feedback, several peer assessments per assignment are needed in addition to a training exercise, and students sometimes complain about workload. To reduce the effort, Kulkarni et al. (2014) experimented with a workflow that uses machine grading as a first step. After training a machine learning algorithm for a given assignment, assignments are scored by the algorithm. The less confident the algorithm is in its score, the more students are assigned to grade the assignment, but high-confidence assignments may need only one peer grader. This step was found to successfully reduce the amount of feedback needed to be done with a moderate decrease in grading performance. That said, the algorithm did require the instructors to mark up 500 sample assignments, and there is room for improvement in the algorithm in other ways, since only a first pass at NLP techniques was used to date.

Nonetheless, mixing machine and peer grading is a promising technique to explore, as it has been found to be useful in other contexts (Nguyen and Litman, 2014; Kukich, 2000).

5 Are You a FakeBot?

Why is the completion rate of MOOCs so low? This question vexes proponents and opponents of MOOCs alike. Counting the window shopping enrollees of a MOOC who do not complete a course is akin to counting everyone who visits a college campus as a failed graduate of that university; many people are just checking the course out (Jordan, 2014). That said, although the anytime, anywhere aspect of online courses works well for many busy professionals who are self-directed, research shows that most people need to learn in an environment that includes interacting with other people.

Learning with others can refer to instructors and tutors, and online tutoring systems have had success comparable to that of human tutors in some cases (VanLehn, 2011; Aleven et al., 2004). But another important component of learning with others refers to learning with other students. Literally hundreds of research papers show that an effective way to help students learn is to have them talk together in small groups, called structured peer learning, collaborative learning, or cooperative learning (Johnson et al., 1991; Lord, 1998). In the classroom, this consists of activities in which students confer in small groups to discuss conceptual questions and to engage in problem-solving. Studies and meta-analyses show the significant pedagogical benefit of peer learning including improved critical thinking skills, retention of learned information, interest in subject matter, and class morale (Hake, 1998; Millis and Cottell, 1998; Springer et al., 1999; Smith et al., 2009; Deslauriers et al., 2011). Even studies of intelligent tutoring systems find it hard to do better than just having students discuss homework problems in a structured setting online (Kumar et al., 2007).

The reasons for the success of peer learning include: students are at similar levels of understanding that experts can no longer relate to well, people learn material better when they have to explain it to others, and identify the gaps in their current understanding, and the techniques of structured peer learning introduce activities and incentives to help students help one another.

S2	I think E is the right answer
S1	Hi, I think E is right, too
S3	Hi! This seems to be a nurture vs nature question.
S3	Can scent be learned, or only at birth?
S2	Yeah, but answer A supports the author's conclusion
S1	I felt that about A too
S2	But the question was, which statement would weaken the author's conclusion
S3	So I choose A, showing that scent can be learned at not only AT BIRTH.
S2	That's why I think E is right
S3	Are you real, or fake?
S2	real
S1	I didn't think that b or d had anything to do with the statement
S3	Actually what you said makes sense.
S1	So, do we all agree that E was the correct answer?
S2	I think so, yes.
S3	But I'm sticking with A since "no other water could stimulate olfactory sites" abd I suggests that other water could be detected.
S3	*and
S1	I thought about c for awhile but it didn't really seem to have anything to do with the topic of scent
S3	It has to be A or E. Other ones don't have anything do do with the question.
S2	but that "no other water" thing applies equally well to E
S3	E is still about spawing ground water, I think. this is a confusing question.
S1	I thought E contradicted the statement the most
S2	me too
S3	I loving hits with other mturkers

Table 1: Transcript of a conversation among three crowdworkers who discussed the options for a multiple choice question for a GMAT logical reasoning task. Note the meta-discussion about the prevalence of robots on the crowdsourcing platform.

In our MOOCChat research, we were interested in bringing structured peer learning into the MOOC setting. We first tried out the idea on a crowdsourcing platform (Coetzee et al., 2015), showing that when groups of 3 workers discussed challenging problems together, and especially if they were incentivized to help each other arrive at the correct answer, they achieved better results than working alone. (A sample conversation is shown in Table 1.) We also found that providing a *mini-lesson* in which workers consider the principles underlying the tested concept and justify their answers leads to further improvements, and combining the mini-lesson with the discussion of the corresponding multiple-choice question in a group of 3 leads to significant improvements on that question. Crowd workers also expressed positive subjective responses to the peer interactions, suggesting that discussions can improve morale in remote work or learning settings.

When we tested the synchronous small-group discussions in a live MOOC we found that, for those students that were successfully placed into a group of 3 for discussion, they were quite positive about the experience (Lim et al., 2014). However, there are significant challenges in getting students to coordinate synchronously in very large low-cost courses (Kotturi et al., 2015).

There is much NLP research to be done to enhance the online dialogues that are associated with student discussion text beyond the traditional role of intelligent tutoring systems. One idea is to monitor discussions in real time and try to shape the way the group works together (Tausczik and Pennebaker, 2013). Another idea is to automatically assess if students are discussing content at appropriate levels on Bloom's taxonomy of educational objectives (Krathwohl, 2002).

In our MOOCChat work with triad discussions we observed that more workers will change their answer from an incorrect to a correct one if at least one member of the group starts out correct than if no one is correct initially (Hearst et al., 2015). We also noticed that if all group members start out with the same answer — right or wrong — no one is likely to change their answer in any direction. This behavior pattern suggests an interesting idea for large scale online group discussions that are not feasible in in-person environments: dynamically assign students to groups depending on what their initial answers to questions are, and dynamically regroup students according to the misconceptions and correct conceptions they have. Rather than building an intelligent tutoring system to prompt students with just the right statement at just the right time, a more successful strategy might be to mix students with other poeple who for that particular discussion point have the just the right level of conceptual understanding to move the group forward.

6 Conclusions

In this paper I am suggesting inverting the standard mode of our field from that of processing, correcting, identifying, and generating aspects of language to one of recognizing what a person is doing with language: NLP algorithms as coaches rather than critics. I have outlined a number of specific suggestions for research that are currently outside the mainstream of NLP research but which pose challenges that I think some of my colleagues will find interesting. Among these are text analyzers that explain what is wrong with an essay at the clause, sentence, discourse level as the student writes it, promoting mastery learning by generating unlimited practice problems, with answers, in a form that makes practice fun, and using NLP to improve the manner in which peers learning takes place online. The field of learning and education is being disrupted, and NLP researchers should be helping push the frontiers.

Acknowledgements I thank the ACL program chairs Michael Strube and Chengqing Zong for inviting me to write this paper and keynote talk, Lucy Vanderwende for suggested references, and colleagues at NAACL 2015 for discussing these ideas with me. This research is supported in part by a Google Social Interactions Grant.

References

Vincent Aleven, Amy Ogan, Octav Popescu, Cristen Torrey, and Kenneth Koedinger. 2004. Evaluating the effectiveness of a tutorial dialogue system for self-explanation. In *Intelligent tutoring systems*, pages 443–454. Springer.

Divya Anand, Shreyas, Sonali Sharma, Victor Starostenko, Ashley DeSouza, Kimiko Ryokai, and Marti A. Hearst. 2015. *Wordcraft: Playing with Sentence Structure*. Under review.

Lars Borin. 2002. *What have you done for me lately? The fickle alignment of NLP and CALL.* Reports from Uppsala Learning Lab.

Michael Brooks, Sumit Basu, Charles Jacobs, and Lucy Vanderwende. 2014. Divide and correct: Using clusters to grade short answers at scale. In *Proceedings of the first ACM conference on Learning@Scale*, pages 89–98. ACM.

D Coetzee, Seongtaek Lim, Armando Fox, Bjorn Hartmann, and Marti A Hearst. 2015. Structuring interactions for large-scale synchronous peer learning. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 1139–1152. ACM.

Robert Dale and Adam Kilgarriff. 2011. Helping our own: The HOO 2011 pilot shared task. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 242–249. Association for Computational Linguistics.

Robert Dale, Ilya Anisimoff, and George Narroway. 2012. HOO 2012: A report on the preposition and determiner error correction shared task. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 54–62. Association for Computational Linguistics.

Louis Deslauriers, Ellen Schelew, and Carl Wieman. 2011. Improved learning in a large-enrollment physics class. *Science*, 332(6031):862–864.

Darren Edge, Elly Searle, Kevin Chiu, Jing Zhao, and James A Landay. 2011. Micromandarin: mobile language learning in context. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3169–3178. ACM.

Darren Edge, Kai-Yin Cheng, Michael Whitney, Yao Qian, Zhijie Yan, and Frank Soong. 2012. Tip tap tones: mobile microtraining of mandarin sounds. In *Proceedings of the 14th International Conference on Human-Computer Interaction with Mobile Devices and Services*, pages 427–430. ACM.

Sumit Gulwani. 2014. Example-based learning in computer-aided stem education. *Communications of the ACM*, 57(8):70–80.

Richard R Hake. 1998. Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics*, 66(1):64–74.

Andrew Head, Yi Xu, and Jingtao Wang. 2014. Tonewars: Connecting language learners and native speakers through collaborative mobile games. In *Intelligent Tutoring Systems*, pages 368–377. Springer.

Marti A Hearst, Armando Fox, D Coetzee, and Bjoern Hartmann. 2015. All it takes is one: Evidence for a strategy for seeding large scale peer learning interactions. In *Proceedings of the Second (2015) ACM Conference on Learning@Scale*, pages 381–383. ACM.

George Heidorn. 2000. Intelligent writing assistance. *Handbook of Natural Language Processing*, pages 181–207.

Michael Heilman and Nitin Madnani. 2015. The impact of training data on automated short answer scoring performance. In *Proceedings of the Tenth Workshop on Building Educational Applications Using NLP*. Association for Computational Linguistics.

Catherine M Hicks, C Ailie Fraser, Purvi Desai, and Scott Klemmer. 2015. Do numeric ratings impact peer reviewers? In *Proceedings of the Second (2015) ACM Conference on Learning@ Scale*, pages 359–362. ACM.

Lynette Hirschman, Eric Breck, Marc Light, John D Burger, and Lisa Ferro. 2000. Automated grading of short-answer tests. *Intelligent Systems and their Applications, IEEE*, 15(5):22–37.

David W Johnson, Roger T Johnson, and Karl Aldrich Smith. 1991. *Active learning: Cooperation in the college classroom*. Interaction Book Company Edina, MN.

Katy Jordan. 2014. Initial trends in enrolment and completion of massive open online courses. *The International Review Of Research In Open And Distributed Learning*, 15(1).

Salman Khan. 2012. *The One World Schoolhouse: Education Reimagined*. Twelve.

Yasmine Kotturi, Chinmay Kulkarni, Michael S Bernstein, and Scott Klemmer. 2015. Structure and messaging techniques for online peer learning systems that increase stickiness. In *Proceedings of the Second (2015) ACM Conference on Learning@ Scale*, pages 31–38. ACM.

David R Krathwohl. 2002. A revision of Bloom's taxonomy: An overview. *Theory into practice*, 41(4):212–218.

Karen Kukich. 2000. Beyond automated essay scoring. *IEEE Intelligent Systems and their Applications, IEEE*, 15(5):22–27.

Chinmay Kulkarni, Koh Pang Wei, Huy Le, Daniel Chia, Kathryn Papadopoulos, Justin Cheng, Daphne Koller, and Scott R Klemmer. 2013. Peer and self assessment in massive online classes. In *ACM Transactions on Computer Human Interaction (TOCHI)*, volume 20. ACM.

Chinmay E Kulkarni, Richard Socher, Michael S Bernstein, and Scott R Klemmer. 2014. Scaling short-answer grading by combining peer assessment with algorithmic scoring. In *Proceedings of the first ACM conference on Learning@Scale*, pages 99–108. ACM.

Rohit Kumar, Carolyn Penstein Rosé, Yi-Chia Wang, Mahesh Joshi, and Allen Robinson. 2007. Tutorial dialogue as adaptive collaborative learning support. *Frontiers in Artificial Intelligence and Applications*, 158:383.

Thomas K Landauer, Darrell Laham, and Peter W Foltz. 2000. The intelligent essay assessor. *IEEE Intelligent Systems and their Applications, IEEE*, 15(5):22–27.

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

Seongtaek Lim, Derrick Coetzee, Bjoern Hartmann, Armando Fox, and Marti A Hearst. 2014. Initial experiences with small group discussions in moocs. In *Proceedings of the first ACM conference on Learning@Scale*, pages 151–152. ACM.

Yun-En Liu, Christy Ballweber, Eleanor O'rourke, Eric Butler, Phonraphee Thummaphan, and Zoran Popović. 2015. Large-scale educational campaigns. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 22(2):8.

Thomas Lord. 1998. Cooperative learning that really works in biology teaching: using constructivist-based activities to challenge student teams. *The American Biology Teacher*, 60(8):580–588.

Detmar Meurers. 2012. Natural language processing and language learning. In *The Encyclopedia of Applied Linguistics*. Wiley Online Library.

Barbara J Millis and Philip G Cottell. 1998. *Cooperative learning for higher education faculty*. Oryx Press (Phoenix, Ariz.).

Michael Mohler and Rada Mihalcea. 2009. Text-to-text semantic similarity for automatic short answer grading. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 567–575. Association for Computational Linguistics.

Hwee Tou Ng, Siew Mei Wu, Yuanbin Wu, Christian Hadiwinoto, and Joel Tetreault. 2014. The CoNLL-2013 shared task on grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task (CoNLL-2014 Shared Task)*, pages 1–12.

Huy V Nguyen and Diane J Litman. 2014. Improving peer feedback prediction: The sentence level is right. *ACL 2014*, page 99.

Oleksandr Polozov, Eleanor ORourke, Adam M Smith, Luke Zettlemoyer, Sumit Gulwani, and Zoran Popovic. 2015. Personalized mathematical word problem generation. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI 2015)*. To appear.

Camilla Schwind. 1988. Sensitive parsing: error analysis and explanation in an intelligent language tutoring system. In *Proceedings of the 12th conference on Computational Linguistics-Volume 2*, pages 608–613. Association for Computational Linguistics.

Serge Sharoff, Stefania Spina, and Sofie Johansson Kokkinakis. 2014. Introduction to the special issue on resources and tools for language learners. *Language Resources and Evaluation*, 48(1):1–3.

Michelle K Smith, William B Wood, Wendy K Adams, Carl Wieman, Jennifer K Knight, Nancy Guild, and Tin Tin Su. 2009. Why peer discussion improves student performance on in-class concept questions. *Science*, 323(5910):122–124.

Leonard Springer, Mary Elizabeth Stanne, and Samuel S Donovan. 1999. Effects of small-group learning on undergraduates in science, mathematics, engineering, and technology: A meta-analysis. *Review of educational research*, 69(1):21–51.

Yla R Tausczik and James W Pennebaker. 2013. Improving teamwork using real-time language feedback. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 459–468. ACM.

Joel Tetreault, Jill Burstein, and Claudia Leacock. 2015. *Proceedings of the Tenth Workshop on Building Educational Applications Using NLP*. Association for Computational Linguistics.

Kurt VanLehn. 2011. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, 46(4):197–221.

Luis von Ahn. 2013. Duolingo: learn a language for free while helping to translate the web. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces*, pages 1–2. ACM.

Lijuan Wang, Yao Qian, Matthew R Scott, Gang Chen, and Frank K Soong. 2012. Computer-assisted audiovisual language learning (with online video). *Computer*, 45(6):38–47.

Alistair Willis. 2015. Using nlp to support scalable assessment of short free text responses. In *Proceedings of the Tenth Workshop on Building Educational Applications Using NLP*. Association for Computational Linguistics.