Feedback Strategies for Form and Meaning in a Real-life Language Tutoring System

Ramon Ziai Björn Rudzewitz
Kordula De Kuthy Florian Nuxoll Detmar Meurers

Collaborative Research Center 833
Department of Linguistics, ICALL Research Group*
LEAD Graduate School & Research Network
University of Tübingen

Abstract

We describe ongoing work on an English language tutoring system currently being used as part of regular instruction in twelve German high school classes. In contrast to the traditional ICALL system approach analyzing learner language, we build on the approach of Rudzewitz et al. (2018) to generate variants of target answers based on task and target language models and combine this offline step with an online process flexibly matching learner answers with these variants. We extend the approach by advancing the search engine used in the online step to return more relevant results. Then we extend the approach to meaning-focused feedback, showing how it can be realized in the system in addition to the form-focused feedback. We conclude with an outlook on an intervention study we have designed to evaluate the system.

1 Introduction

Second Language Acquisition (SLA) has long recognized the need for immediate feedback on learner production (Mackey, 2006). However, in real-life classrooms, there is limited opportunity for such immediate feedback if every student is to be considered according to her needs.

Intelligent Language Tutoring Systems make it possible to address this shortcoming since they offer the possibility of automated, immediate feedback while the learner is working on the task, and many students can use the system at the same time whenever they want to, whereas opportunities for interaction with a teacher or other tutor are much more limited.

However, in order to provide accurate, helpful feedback, the erroneous forms produced by learners need to be characterized. If one analyzes learner language directly, one runs into the problem that state-of-the-art NLP is not equipped to deal with non-standard language in a way that supports fine-grained feedback. This is not surprising given that the linguistic categories system was developed for well-formed, native language, thus NLP tools generally treat the analysis of learner language as a robustness problem, covering up the type of deviation or error that the learner produced instead of characterizing it (Díaz Negrillo et al., 2010; Meurers and Dickinson, 2017). As an example, consider that a standard POS tagger would typically assign the tag `VBD` to the overregularized form *teached* based on the suffix analysis fallback strategy commonly used for unknown words.

If we know what task the learner language was produced for, this challenge can be addressed to some degree: instead of analyzing the learner productions directly, one can start out from the expected target forms and systematically transform them into well-formed and ill-formed variations of the target (Rudzewitz et al., 2018).

In this paper, we expand on that idea and present feedback strategies supporting both form- and meaning-oriented tasks. After reviewing the process responsible for generating the well-formed and ill-formed variation, we zoom in on the search process executed at feedback time, outlining how standard search engine technology was adapted to serve the needs of a language tutoring system. We show how the same basis of generated variation can support meaning-oriented feedback, using an alignment-based approach inspired by research on Short Answer Assessment (Meurers et al., 2011). We demonstrate this with feedback for reading and listening comprehension where the learner is pointed to the relevant source of information in

* http://icall-research.de

Ramon Ziai, Bjoern Rudzewitz, Kordula De Kuthy, Florian Nuxoll and Detmar Meurers 2018. Feedback strategies for form and meaning in a real-life language tutoring system. *Proceedings of the 7th Workshop on NLP for Computer Assisted Language Learning at SLTC 2018 (NLP4CALL 2018).* Linköping Electronic Conference Proceedings 152: 91–98.

the task material. We conclude with an outlook on the design of an intervention study we are currently running in twelve 7th grade classrooms in Germany.

2 System Setup

The feedback strategies discussed in this article are implemented as part of a web-based online workbook, the FeedBook (Rudzewitz et al., 2018; Meurers et al., 2018). The foreign language tutoring system is an adaptation of a paper workbook for a 7th grade English textbook approved for use in German high schools.

Figure 1 provides an authentic example of a student solution to an exercise in the printed workbook on the use of type II conditionals. For such paper-based exercises, feedback is typically given in a delayed fashion by the teacher, when discussing the exercise summarily in class or sometimes by returning marked-up exercise sheets, not while the student is actually thinking about and working on the task.

In contrast, the system we describe provides an interface for students to select and interactively work on exercises. For exercises that aim at teaching grammar topics, students receive automatic, immediate feedback by the system informing them whether their answer is correct (via a green check mark) or *why* their answer is incorrect (via red color, highlighting of the error span, and a meta-linguistic feedback message). In fact, rather than pointing out the error as such, we instead formulate scaffolding feedback messages designed to guide the learner towards the solution, without giving it away.

The process of entering an answer and receiving feedback can be repeated, incrementally leading the student to the correct answer. If there are multiple errors in a learner response, the system presents the feedback one at a time. Figure 2 shows the same learner production we saw in Figure 1 together with the interactive feedback immediately provided by the system after this is typed in.

Students can save and resume work, interact with the system to receive automatic feedback and revise their answers, and eventually submit their final solutions to the teacher. In case the answers in a given exercise are all correct, the system grades the submission automatically, without requiring teacher interaction.

For those answers that are not correct with respect to a given target answer, the teacher can manually annotate the learner answer with feedback parallel to the traditional mark-up process known from printed workbooks. Any such manual feedback is saved in a feedback memory and suggested automatically to the teacher in case the form occurs in another learner response to this exercise.

The system also provides students with automatic, immediate feedback for many exercise types, where they traditionally would either not receive it or only after long delay resulting from collecting and manually marking up homework assignments. From the teacher's perspective, the system relieves them from very repetitive and time-consuming work. The exercises are embedded in a full web application with a messaging system for communication, a profile management including e-mail settings, tutorials for using the system, classroom management, and various functions orthogonal to the NLP-related issues.

3 Hypothesis Generation Revisited

The generation of well-formed and ill-formed answers expected for a given exercise builds on the generation framework proposed by Rudzewitz et al. (2018) which generate variants of target answers for each task that one wants to provide feedback for.

The crucial components of the framework are i) a set of rules organized in layers that transform one variant to another variant, introducing one change at a time, ii) a common representation format for adding, removing and querying units of linguistic analysis (the CAS, see Götz and Suhre 2004), and iii) a breadth-first search algorithm that traverses the rule layers, applying rules and passing the output variants of rules to rules in the next layer along with their linguistic analysis.

The setup consists of four layers: in the first layer normalizations like contractions are performed. In the second layer transformations are conducted that yield linguistically well-formed, but task-inappropriate forms like tense changes. As the next step, the third layer introduces changes that result in morphologically ill-formed answers, for example regular endings for irregular verbs in the simple past. Finally, the fourth layer rejoins and normalizes different generated variants. Not every layer introduces new diagnoses: for

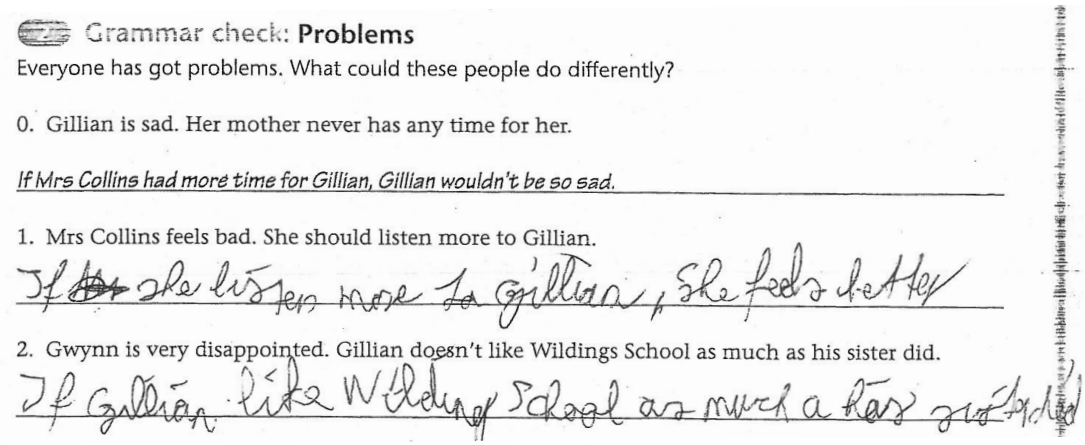

Figure 1: Traditional paper-based exercise

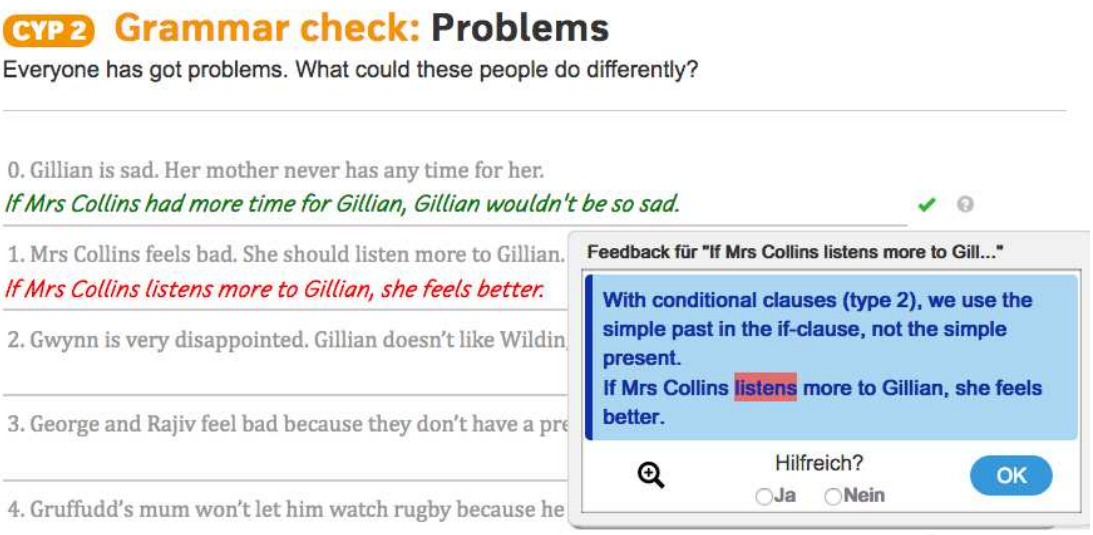

Figure 2: Interactive exercise with form-oriented feedback

the normalization rules, the previous diagnoses are passed on. At each point, the current variant and corresponding analysis is saved so they can be used later for feedback. The system (at the time of publication) generated 95.386 distinct hypotheses for 3.211 target answers.

Table 1 provides some example derivations that result from rule interactions.

4 The Search Mechanism

Given the generation approach outlined in the previous section, it should come as no surprise that especially for short-answer tasks such as the one in Figure 2, the number of generated variants can get very large. This is especially true for items with multiple target answers given that a separate calculation is done for every target answer.

When a learner uses the system and triggers the feedback mechanism for a given item, it is necessary to compare the learner answer to the relevant pre-stored generated variants, determine whether the student made one of the errors present (and thus known to the system) in the variant, and if so, provide feedback. Since it is infeasible to traverse and compare all variants, Rudzewitz et al. (2018) use the search engine framework Lucene[1] to efficiently index and query the stored variants. Every variant is treated like a document indexed by Lucene.

In examining the feedback behavior of such a system, we noticed that Lucene did not always return the most relevant variant for a given learner answer and task. Looking further into this issue, we discovered that this behavior was due to the term weighting scheme used by Lucene and other search engines, known as TF-IDF (Salton and McGill, 1983). TF-IDF works by balancing the frequency of a word in a given document (TF) against the inverse frequency of the word occurring over the whole set of documents (IDF), resulting in low values for very frequent words, and high values for topic-specific words only occurring in few documents.

While this is the desired behavior when looking for specific content, it is not suitable for the present problem of finding relevant variants for learner answers. We therefore modified the approach of Rudzewitz et al. (2018) by i) eliminating the IDF part of the weighting scheme, and ii) introducing task-specific term weighting into the search.

In order to realize the latter, we draw on information gathered during the generation process. We always store the transformation result r of a rule application, i.e., the part of a variant which was changed by the rule. So the set of all transformation results R is known before the learner interacts with the system. We can thus look for instances of each $r \in R$ of a given task and item (such as the incorrect tense forms shown in Table 1) in the learner answer and assign a higher weight for parts of the learner answer that match r. The weighting is implemented using a Lucene feature called "query boosting", which allows for assignment of different weights to sub-strings of the query. We use the same weight for all matches (currently 5.0), whereas the non-matched answer parts receive the standard weight of 1.0.

As a result of this modification, the system is able to give more task-relevant feedback for the learner answer in Figure 2 despite a low token overlap of the learner answer with the target answer. In order to also obtain a quantitative result, we ran the new search mechanism against the same data used for coverage testing in Rudzewitz et al. (2018): we observed a 6% increase in types of answers covered by construction-specific feedback (16.9% / 1085 instances vs. 10.9% / 696 instances) as a consequence of the search mechanism introduced above.

5 Meaning-oriented Feedback

Depending on the nature of the exercise, it is essential to draw the learners' attention not just to forms but also to content-related misconceptions. Indeed, for meaning-based exercises such as listening- or reading comprehension, feedback on meaning should take priority over feedback on form. The strategies needed to detect such errors are very different from the ones used for the form-oriented feedback described so far. In contrast to analyzing or generating variation in form, one needs to abstract over it and recognize meaning equivalence of different forms. A learner answer can then be accepted as correct whenever meaning equivalence has been established between it and the target answer.

There is a vast body of work on automated short-answer grading (see Burrows et al. 2015 for an overview), but the overwhelming majority of

[1] https://lucene.apache.org/core/

target	layer 1	layer 2	layer 3
are you doing	are you doing	are you doing	are you doing
	were you doing	were you do	was you do
	have you been doing	have you been do	have you been dos
	had you been doing	had you been do	had you been dos
	will you do	are you do	will you dos
	did you do	...	did you dos
	...		are you dos
			was you doing
			is you dos
			is you doing
			...
friendlier	friendlier	friendlier	friendlier
	more friendly	more friendlier	most friendlier
	friendlyer	more friendlyer	most friendlyer
	...		friendliest
		...	friendlyest
			...

Table 1: Examples for generated answer variants

work only lends itself to the task of holistically scoring learner answers, not detecting the type of divergence from target answers. We chose to adapt the alignment-based CoMiC system (Meurers et al., 2011) to our needs. Instead of classifying learner productions, we use the alignment information from CoMiC as evidence for equivalence or divergence (e.g., missing information) of the learner answer from the target answer.

Given the means of detecting meaning errors, the question arises how to point the student in the right direction. Since it is pedagogically not acceptable to reveal (parts of) the correct answer, an alternate means of scaffolding for meaning-oriented exercises such as reading and listening comprehension is needed. How can this be done?

Our general approach is to draw the learner's attention to relevant parts of the task context. This can be a part of the reading text or listening clip, the question being asked, or the instruction text. Figure 3 shows feedback on a learner answer with missing information. The system reacts to the problem by visually highlighting the relevant part of the reading text, pointing the learner into the direction of the correct answer.

For listening comprehension exercises, the overall strategy is the same, but instead of highlighting or displaying text, we provide an excerpt of the corresponding audio clip that contains the information necessary for answering the given question. Figure 4 illustrates an example for such feedback. Since the current number of suitable tasks in FeedBook is limited, a teacher from the project team manually specified the relevant part of the task context for each task. In the future, we plan to automatically identify these information sources in reading texts or transcripts of listening texts. Furthermore, we are in the process of compiling a test suite for meaning-oriented feedback in order to quantitatively evaluate our approach.

6 Summary

We presented extensions to the language tutoring system FeedBook currently in use in English 7th grade classrooms. The extensions are i) a task-based optimization of the search strategy necessary when comparing learner answers to prestored variants and ii) the addition of meaning-oriented scaffolding feedback for reading and listening activities. We demonstrated both extensions by example. The first extension shows that if the task is known and target answers exist, it is possible to give accurate feedback on learner language without having to directly process it. The second extension makes it possible to give helpful, pedagogically sound scaffolding feedback on meaning-oriented tasks.

7 Outlook: Towards a Large-Scale Intervention Study

Moving forward, it will be necessary to evaluate the effectiveness of the system in terms of learning outcomes. Very few ICALL systems have been evaluated in real-life formal learning contexts (for some notable exceptions, cf. Nagata, 1996; Heift, 2004, 2010; Choi, 2016), let alone in terms of standards for intervention studies established in psychology and empirical educational science. However, in order to raise awareness for and show the impact of ICALL systems, it arguably is crucial

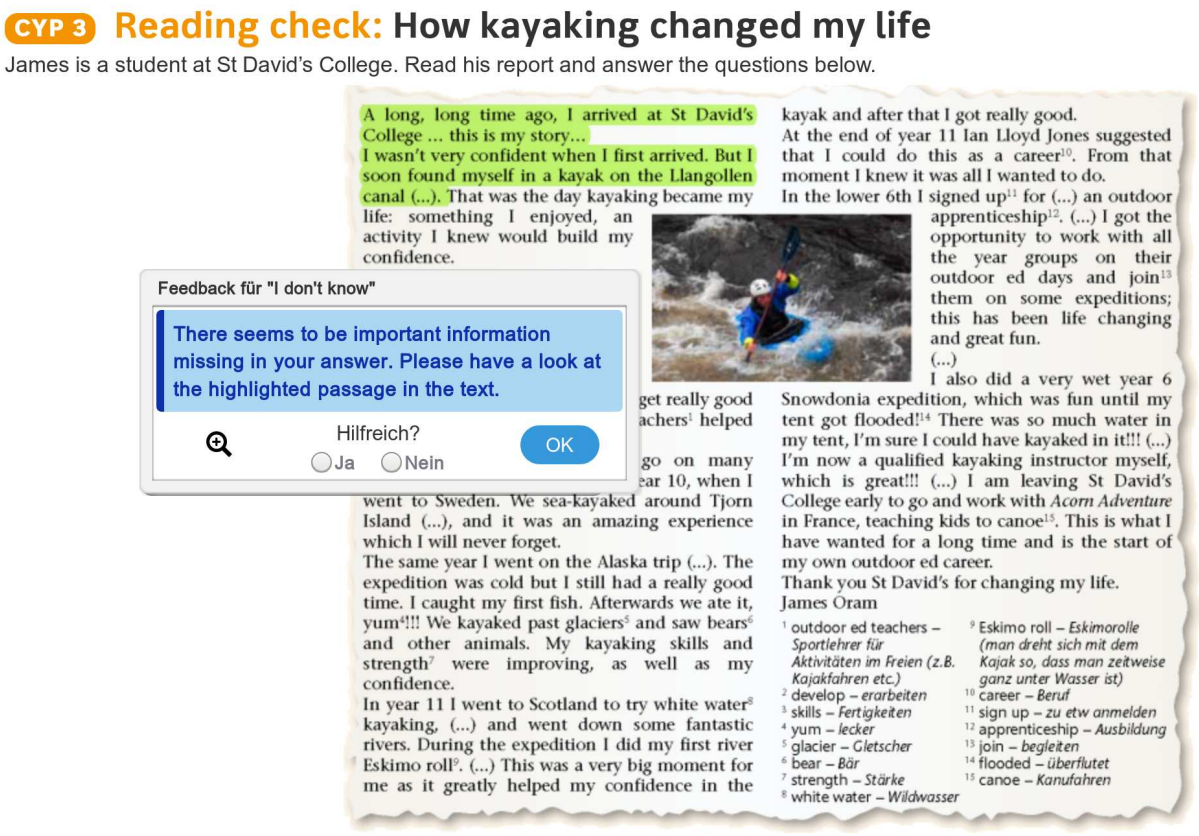

Figure 3: Meaning-oriented feedback for reading comprehension exercise

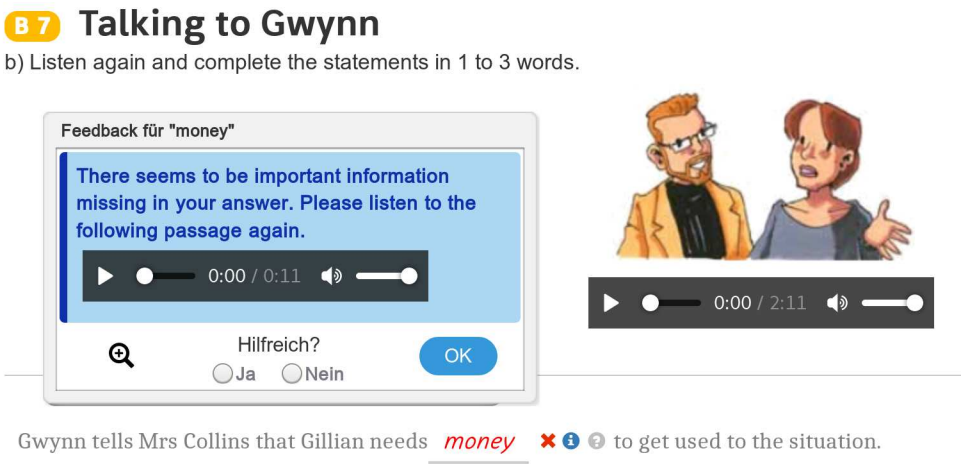

Figure 4: Meaning-oriented feedback for listening comprehension exercise

to provide large-scale evaluation in terms of externally established measures of learning outcomes. In our case, we want to measure the impact of interactive feedback on the individual learning outcomes of 7th grade school children.

We have set up a randomized controlled field study that compares two groups of students receiving immediate feedback on different grammatical constructions throughout the current school year. The variables that are relevant to control in such a context include: a) the learners' language proficiency, b) individual differences in aptitude/cognition, c) motivational factors, and, last but far from least, d) the teacher, known to have the strongest influence on learning outcome in classrooms.

For a), we plan to administer both a C-Test measuring general language proficiency as well as a construction-specific grammar test geared towards testing grammar topics that are part of the 7th grade English curriculum. When piloting the grammar test, we observed that conducting a systematic pre-test of all constructions at the beginning of the school year, before the students have covered these constructions in class, is very time consuming and leads to significant student frustration. Students are not used to being tested on material they have not systematically covered in school yet. So for the main study, we are distributing the pre-tests of the grammatical constructions throughout the school year to just before the specific construction is being covered in class.

In order to control for b), we will employ established individual difference tests such as MLAT-5 (Carroll and Sapon, 1959) to determine fixed traits of learners, such as working memory capacity. For motivation and other background traits (c), we will use a questionnaire where students answer a range of questions on the subject they learn, the languages they speak, and other relevant information. Originally, we had planned to administer all these tests using our web-based platform. To ensure that these tests are conducted systematically, this is supposed to happen in class.

It turns out, however, that in the current state of the German secondary school system, the overhead of scheduling classes in computer rooms providing a sufficient number of computers that are functional and connected to the Internet is a significant burden for teachers. Conducting tests on paper, on the other hand, means having to manually enter the data later, which for studies of this size is very work intensive and error prone. For some of the individual difference tests, it is possible, though, to let students complete them at home using the digital device they also use to access the tutoring system. In pilot testing some tests in such a way outside of class, we found that in such a setting it is very difficult to ensure that all students actually complete the tests. To enforce completion, in the main study we are only making the interactive online exercises for the next chapter available in the tutoring system once the tests scheduled at that point have been completed by a student.

To account for the teacher factor (d), the intervention study uses within-class randomization. We divided the grammar topics in the curriculum into two groups and assign students randomly to one of these groups. Students get immediate system feedback on the constructions assigned to their group, while not receiving automated feedback on the other grammar topics. Both groups thus receive feedback from the system, but systematically for different constructions. If the interactive feedback is effective, the two student groups should differ in their performance on the different grammatical construction and general language proficiency posttest. Except for the presumably stable traits, such as working memory and the background and motivation questionnaires, all tests are administered following a pre-/posttest design.

In addition to the twelve test classes with within-class randomization, we also recruited a separate class as a business-as-usual control, where the traditional paper workbook is used and only the tests are administered. We intentionally did not make the comparison with business-as-usual the main focus of our study since we want to determine the effect of interactive scaffolding feedback on learning, not the well-known newness effect of using a web-based computer system in comparison to a paper-based workbook.

Acknowledgements

We are grateful to the high school students, parents and teachers using the FeedBook system and providing much useful feedback. We would also like to thank the reviewers for their detailed and helpful comments.

References

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(1):60–117.

John B. Carroll and Stanley M. Sapon. 1959. *Modern language aptitude test*. Psychological Corporation, San Antonio, TX, US.

Inn-Chull Choi. 2016. Efficacy of an ICALL tutoring system and process-oriented corrective feedback. *Computer Assisted Language Learning*, 29(2):334–364.

Ana Díaz Negrillo, Detmar Meurers, Salvador Valera, and Holger Wunsch. 2010. Towards interlanguage POS annotation for effective learner corpora in SLA and FLT. *Language Forum*, 36(1–2):139–154.

Thilo Götz and Oliver Suhre. 2004. Design and implementation of the uima common analysis system. *IBM Systems Journal*, 43(3):476–489.

Trude Heift. 2004. Corrective feedback and learner uptake in call. *ReCALL*, 16(2):416–431.

Trude Heift. 2010. Prompting in CALL: A longitudinal study of learner uptake. *Modern Language Journal*, 94(2):198–216.

Alison Mackey. 2006. Feedback, noticing and instructed second language learning. *Applied Linguistics*, 27(3):405–430.

Detmar Meurers and Markus Dickinson. 2017. Evidence and interpretation in language learning research: Opportunities for collaboration with computational linguistics. *Language Learning*, 67(2).

Detmar Meurers, Kordula De Kuthy, Verena Möller, Florian Nuxoll, Björn Rudzewitz, and Ramon Ziai. 2018. Digitale Differenzierung benötigt Informationen zu Sprache, Aufgabe und Lerner. Zur Generierung von individuellem Feedback in einem interaktiven Arbeitsheft [Digitial differentiation requires information on language, task, and learner. On the generation of individual feedback in an interactive workbook]. *FLuL – Fremdsprachen Lehren und Lernen*, 47(2):64–82.

Detmar Meurers, Ramon Ziai, Niels Ott, and Stacey 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.

Noriko Nagata. 1996. Computer vs. workbook instruction in second language acquistion. *CALICO Journal*, 14(1):53–75.

Björn Rudzewitz, Ramon Ziai, Kordula De Kuthy, Verena Möller, Florian Nuxoll, and Detmar Meurers. 2018. Generating feedback for English foreign language exercises. In *Proceedings of the 13th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, pages 127–136. ACL.

Gerard Salton and Michael J. McGill. 1983. *Introduction to modern information retrieval*. McGraw-Hill, New York.