International Society of the Learning Sciences

How to Use Theory to Implement Natural Language Processing for Peer-Feedback

Martin Greisel*, University of Augsburg, martin.greisel@uni-a.de
Elisabeth Bauer*, Ludwig-Maximilians-Universität in Munich, elisabeth.bauer@psy.lmu.de
Ilia Kuznetsov, Technical University of Darmstadt, kuznetsov@ukp.informatik.tu-darmstadt.de
Markus Berndt, University Hospital, LMU Munich, markus.berndt@med.uni-muenchen.de
Markus Dresel, University of Augsburg, markus.dresel@uni-a.de
Martin Fischer, University Hospital, LMU Munich, martin.fischer@med.uni-muenchen.de
Ingo Kollar, University of Augsburg, ingo.kollar@uni-a.de
Frank Fischer, Ludwig-Maximilians-Universität in Munich, frank.fischer@psy.lmu.de
* shared first authorship

Abstract: Whenever learners produce text, natural language processing (NLP) has great potential to improve learning. Theories from learning sciences should guide the implementation of NLP into concrete learning scenarios. However, theoretical concepts are much more abstract than the targets and inputs NLP can work with. Therefore, a process is needed which translates theory into NLP tasks. As such a process is missing, we propose a terminological and procedural scheme which researchers and practitioners can employ to develop NLP-based adaptive support measures for learning processes. It defines a sequence of leverage points, support measures, adaptation targets, automation goals, data, prediction targets, input, intrinsic metrics, NLP model, and extrinsic metrics. To illustrate it, we apply it to peer-feedback as a use case.

Problem statement

Recent developments in artificial intelligence (AI) promise to foster various learning processes (Dawson et al., 2019). Yet, to do so, theories of learning are indispensable (Wise & Shaffer, 2015). For example, peer-feedback is assumed to boost learning because, among other reasons, it provides feedback which can be easier to understand than teacher feedback and constitutes an additional learning opportunity (Li et al., 2020). However, for this purpose, learners should provide high-quality feedback (Patchan et al., 2016), but not all learners have the necessary prior knowledge and skills. Consequently, AI might support the learners when composing their feedback. As feedback often is written text, natural language processing (NLP) is the most relevant field of AI for this task. Yet, applying NLP to peer-feedback is complex: NLP and learning sciences have their own terminologies and approaches to conceptualize phenomena. Hence, mapping the theoretical learning process to NLP becomes a complex endeavor of synchronizing terminology, goals, and procedures. A guiding framework that helps researchers and practitioners through this development process is missing. For this reason, in this conceptual paper, we propose a terminological and procedural scheme to guide the development of NLP support measures and exemplify it for peer-feedback.

Crossing disciplinary boundaries

When investigating learning processes employing NLP methods, the usual textual data investigated is much more complex than traditional quantitative data. It might seem tempting to engage in purely exploratory data mining. However, it has been repeatedly argued that theory is needed to inform data analysis, especially if the amount of data is large. For example, Wise and Shaffer (2015, p. 9) argue that, among other functions, "theory gives a researcher guidance about which variables to include in a model", "[…] about how to make results actionable", and "[how to] generalize results to other contexts and populations". The Learning Sciences provide such theories. However, concepts from the Learning Sciences are not immediately usable to build NLP models to address them. For example, empirical findings might state that the quality of a peer-feedback message is crucial to ensure learning success. Learning scenario designers might then ask how NLP should be implemented to foster feedback quality. Obviously, this question cannot be answered immediately. First, we need to answer more specific questions: By which means can peer-feedback quality be improved? How can these means be operationalized in textual data? Which goal should NLP achieve? For which instructional purpose? Which NLP task architectures are most suitable for this? We designed our scheme below to guide researchers in answering these questions for their specific purposes. To conclude, learning sciences provide theory about learning processes that can be used to identify which processes should be addressed by NLP to maximize positive impact on learning outcomes.

Use case example: Outline of a peer-feedback theory

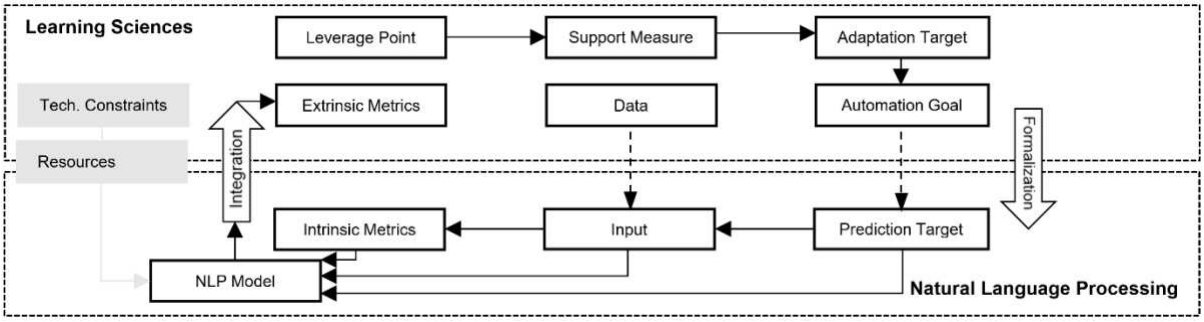
To illustrate how NLP-based support might be developed for a given learning process, we use the following outline of a peer-feedback theory as an example. A common way to employ peer-feedback in teaching roughly involves the subsequent steps: An instructor provides learners with a task which each learner works on individually to produce an initial solution. Then, learners' initial solutions are distributed among the learning group such that each learner receives initial solutions from, for example, two peers. Now, the learner takes the role of a reviewer and examines their peers' initial solutions to compose a feedback message for each peer. In the next step, each learner receives these feedback messages from their peers and, ideally, uses them to revise their initial solution. Obviously, producing a feedback message is crucial in peer-feedback; hence, we use this activity as an example in the following.

A scheme for designing NLP-based adaptive peer-feedback support

NLP is the primary candidate for automating and augmenting adaptive support for peer-feedback scenarios as products such as the feedback message most often constitute textual data. To connect theory with NLP, our terminological and procedural scheme (Figure 1) links concepts from the Learning Sciences with concepts from NLP to facilitate describing, investigating, and designing NLP-based adaptive measures for supporting peer-feedback. Next, we will explain this scheme, starting in the upper left corner (terms from Figure 1 are italicized).

Figure 1

Terminological and Procedural Scheme for Designing NLP-Based Adaptive Support Measures



To be effective, learning support needs to target relevant *leverage points* in the peer-feedback process. Leverage points are factors and instances in the peer-feedback process that can either facilitate or impair the learning process and outcomes. In the current case, the quality of the feedback message determines its effectiveness for facilitating learning processes and outcomes. Patchan et al. (2016) found that high-quality peer-feedback for essay writing not only points out mistakes regarding the content but also refers to high-level aspects of text structure and argumentation as opposed to only low-level aspects such as grammar and spelling. Additionally, specifying where exactly in the essay changes need to be made helped the feedback recipients to implement a feedback comment. Similarly, Wu and Schunn (2020) found that specifying and explaining a problem in the initial solution increases the implementation rate of the respective feedback comment, besides the comment's content quality. Therefore, a central leverage point for increasing the effectiveness of peer-feedback processes consists in advancing the quality of the peer-feedback messages regarding the relevance and accuracy of the feedback content and the degree of helpful elaboration (location or identification and explanation).

Support measures target such leverage points to increase the effectiveness of a peer-feedback process by providing additional task-relevant information or explanations, or by exerting direct or indirect regulation. For example, Gielen and De Wever (2015) showed that even simple guiding questions helped educational science students to elaborate their feedback more and to provide more negative verifications of their peers' draft of a journal article abstract. Of course, such support measures must be adapted to the task context and content area.

Support measures can be designed such that they vary meaningfully across learners. In this case, they are adjusted to *adaptation targets*, i.e., learner characteristics, processes, or outcomes. Adaptation targets are best measured by observable activities and products, especially if the learner support should be adjusted to a dynamically changing adaptation target such as advancing skills (e.g., Tetzlaff et al., 2021). Possible adaptation targets for adjusting peer-feedback support to would be: (1) the reviewer's feedback skills, indicated by a submitted draft of the current feedback message (e.g., for prompts concerning the feedback elaboration) or feedback messages from prior peer-feedback scenarios (e.g., for prompts on avoiding common structural flaws); (2) the peer's task performance in the reviewed initial solution (e.g., for prompts focusing the feedback); (3) the reviewer's task performance in the initial solution (e.g., for personalizing the degree of feedback support).

Technological advancements in AI can automate routine operations and actions (e.g., Ninaus & Sailer, 2022). To improve the efficacy of peer-feedback messages, two kinds of *automation goals* can be distinguished: supporting individual activities (e.g., prompting to further structure insufficiently structured feedback messages), and modifying individual products (e.g., highlighting important flaws in the to-be-reviewed initial solution). Consequently, automation goals for adaptive peer-feedback support focus on automatically detecting relevant aspects in the feedback messages (in addition to analyzing the prior initial solution texts).

Building on the steps outlined above, a precise *formalization* is possible: The general objective of NLP is to create computational models that make accurate predictions about a *prediction target* based on textual input according to evaluation metrics. This aim encompasses task architectures such as text and text pair classification, span and relation extraction, automatic scoring and ranking, text generation, and cross-document linking.

Modern deep-learning based NLP uses pre-trained Transformer-based language models (e.g., Brown et al., 2020) that, based on large unlabeled textual collections, create generally applicable neural representations of text which can be tailored to particular end tasks. General-purpose NLP frameworks further facilitate the development of new NLP models (e.g., Akbik et al., 2019). NLP has been previously applied to support peer-feedback and related scenarios like scholarly peer review (e.g., Hua et al., 2019; Kuznetsov et al., 2022), argumentative writing and essay grading (e.g., Zhang & Litman, 2021). While evaluation of NLP-based applications in small-scale user studies and limited settings is sometimes performed, no holistic framework for NLP-based peer-feedback support has been proposed to date which limits NLP support adoption at large.

Translating complex real-world scenarios systematically into NLP applications is the focal point of Translational NLP (Newman-Griffis et al., 2021). A generic approach to NLP model development for peer-feedback might proceed as follows. Based on the automation goal and available *data*, an appropriate NLP task architecture is chosen, along with the formal definitions of the prediction target, textual *input*, and evaluation metric. For example, solution grading (automation goal) based on task solution texts (data) can be cast as a score prediction task (architecture), that given plain text (input) predicts a numerical score in a given range (prediction target) with the aim of minimizing the deviation between the true and the predicted score (*intrinsic metric*). Based on the size and availability of the data, *technical constraints* (e.g., hardware limitations) and existing NLP *resources* (e.g., pre-trained language models and corpora), NLP practitioners would collaborate with Learning Scientists to select appropriate NLP methodology for the task. Based on that, an NLP model is created, which in most cases involves labeling a portion of data with the target value, applying a supervised machine learning algorithm that learns to predict this value from inputs, and evaluating the model on a held-out data sample according to the metric.

Once the NLP model reaches adequate performance, it can be *integrated* into the learning environment to support the automation goal, either by fully automating a procedure (e.g., predicting the solution score, i.e., the feedback quality), or by augmenting the learner experience via continuous real-time support or via an adaptive scaffold (e.g., notifying the learner before submitting that the current feedback message is insufficient). A model might perform well intrinsically (e.g., solution score deviation is low) but fail to benefit the chosen support measure. Hence, *extrinsic* evaluation is the crucial step for measuring the adequacy of NLP-based support.

To illustrate this scheme for the feedback production phase: NLP could use the textual data consisting of own and other learners' initial solution, and the feedback message draft as input. Feedback messages are subject to structural analysis cast as span and relation extraction or as text classification task. Prior NLP studies have successfully applied discourse parsing to analyze feedback messages in scholarly peer review. For example, Kuznetsov et al. (2022) propose a corpus of scholarly feedback texts where each sentence is labeled with the pragmatic categories such as strength and weakness (feed-back), todo (feed-forward), and recap (feed-up). Alternatively, Hua et al. (2019) focus on extracting argumentation structures, labeling parts of peer review reports as evaluation, fact, request, etc. Similar efforts exist in the domain of English argumentative essay writing. For example, Nguyen et al. (2016) propose, implement, and deploy an instant feedback system to detect the presence of solutions in peer review texts. Automatic analysis of content level addressed in the feedback texts is not widely studied but might similarly be approached as a text classification or a span extraction NLP task.

In conclusion, we proposed a terminological and procedural scheme that shows how, starting from a Learning Science theory, NLP-based, adaptive support measures might be developed systematically. Though we exemplified this scheme for how to improve peer-feedback quality, we believe that the scheme can also be applied to learning processes beyond peer-feedback. It might inspire researchers and practitioners to develop theory-based, adaptive, and automatized support measures to maximize the effectiveness of their learning scenarios.

References

Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., & Vollgraf, R. (2019). FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American*

Chapter of the Association for Computational Linguistics (pp. 54–59). Minneapolis, Minnesota. Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/N19-4010

Alqassab, M., Strijbos, J. W., & Ufer, S. (2018). Training peer-feedback skills on geometric construction tasks: Role of domain knowledge and peer-feedback levels. *European Journal of Psychology of Education, 33*(1), 11–30. https://doi.org/10.1007/s10212-017-0342-0

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. In *Advances in Neural Information Processing Systems* (pp. 1877–1901). Curran Associates.

Cheng, L., Bing, L., Yu, Q., Lu, W., & Si, L. (2020). APE: Argument pair extraction from peer review and rebuttal via multi-task learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (pp. 7000–7011). Online. Association for Computational Linguistics.

Dawson, S., Joksimovic, S., Poquet, O., & Siemens, G. (2019). Increasing the Impact of Learning Analytics. *Proceedings of the 9th International Conference on Learning Analytics & Knowledge,* (pp. 446–455).

Gielen, M., & De Wever, B. (2015). Structuring peer assessment: Comparing the impact of the degree of structure on peer feedback content. Computers in Human Behavior, 52, 315–325.

Hua, X., Nikolov, M., Badugu, N., & Wang, L. (2019). Argument mining for understanding peer reviews. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 2131–2137). Minneapolis, Minnesota. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1219

Kuznetsov, I., Buchmann, J., Eichler, M., & Gurevych, I. (2022). Revise and resubmit: An intertextual model of text-based collaboration in peer review. *Computational Linguistics, 48*(4), 1–38.

Li, H., Xiong, Y., Hunter, C. V., Guo, X., & Tywoniw, R. (2020). Does peer assessment promote student learning? A meta-analysis. *Assessment & Evaluation in Higher Education, 45*(2), 193–211.

Newman-Griffis, D., Lehman, J. F., Rosé, C., & Hochheiser, H. (2021). Translational NLP: A new paradigm and general principles for natural language processing research. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4125–4138). Online. Association for Computational Linguistics.

Nguyen, H., Xiong, W., & Litman, D. (2016). Instant feedback for increasing the presence of solutions in peer reviews. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations* (pp. 6–10). San Diego, California. Association for Computational Linguistics. https://aclanthology.org/N16-3002.pdf

Ninaus, M., & Sailer, M. (2022). Closing the loop—The human role in artificial intelligence for education. *Frontiers in Psychology, 13*. https://doi.org/10.3389/fpsyg.2022.956798

Patchan, M. M., Schunn, C. D., & Correnti, R. J. (2016). The nature of feedback: How peer feedback features affect students' implementation rate and quality of revisions. *Journal of Educational Psychology, 108*(8), 1098–1120. https://doi.org/10.1037/edu0000103

Tetzlaff, L., Schmiedek, F., & Brod, G. (2021). Developing personalized education: A dynamic framework. *Educational Psychology Review, 33*(3), 863–882. https://doi.org/10.1007/s10648-020-09570-w

Wise, A. F., & Shaffer, D. W. (2015). Why Theory Matters More than Ever in the Age of Big Data. *Journal of Learning Analytics, 2*(2), 5–13. https://doi.org/10.18608/jla.2015.22.2

Wu, Y., & Schunn, C. D. (2020). When peers agree, do students listen? The central role of feedback quality and feedback frequency in determining uptake of feedback. *Contemporary Educational Psychology, 62,* 101897. https://doi.org/10.1016/j.cedpsych.2020.101897

Zhang, H., & Litman, D. (2021). Essay quality signals as weak supervision for source-based essay scoring. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 85–96). Online. Association for Computational Linguistics. https://aclanthology.org/2021.bea-1.9

Acknowledgments

Funding: Stiftung Innovation in der Hochschullehre (project "Facilitating Competence Development through Authentic, Digital, and Feedback-Based Teaching-Learning Scenarios"); German Federal Ministry of Research and Education (FAMULUS-Project 16DHL1040); German Research Foundation (DFG FOR 2385).