Computer-assisted stylistic revision with incomplete and noisy feedback
A pilot study

Christian M. Meyer[†‡] **and Johann Frerik Koch**[†]
[†] Ubiquitous Knowledge Processing (UKP) Lab
Technische Universität Darmstadt, Germany
[‡] Research Training Group AIPHES
Technische Universität Darmstadt, Germany
http://www.ukp.tu-darmstadt.de

Abstract

We investigate how users of intelligent writing assistance tools deal with correct, incorrect, and incomplete feedback. To this end, we conduct an empirical user study around an L1 text revision task for German. Our participants should revise stylistic issues in two given texts using a novel web-based writing environment that highlights potential issues and provides corresponding feedback messages. In comparison to a control group, we find that precision plays a more important role than recall, which confirms previous findings for other languages, issue types, user groups, and experimental setups.

1 Motivation

The importance of well-written texts is striking. Research stalls if scientists cannot understand a paper. Technical systems are hardly usable if their documentation is miserable. Job applications may fail due to the use of inadequate registers in a résumé or cover letter. News articles seem carelessly researched if they are full of spelling errors. Even for apparently informal text types, such as microblog posts, authors have to think about a suitable formulation to convey their message in an adequate way to the desired target audience.

To improve a text, authors typically rely on manually provided feedback from friends, colleagues, or professionals as well as on automatically generated feedback from word processors. Since automatic feedback is much less time-consuming and repeatedly available with practically no waiting time, this solution is very attractive.

However, the natural language processing methods generating this kind of feedback are still prone to many errors. Although human feedback might be erroneous as well, automatic methods yet perform significantly worse. The best submission to the 2014 CoNLL shared task on grammatical error correction (Ng et al., 2014) reaches, for example, only 73 % of the average human performance (Bryant and Ng, 2015). A particular issue with automatic feedback is that answers might be embarrassingly wrong (e.g., WATSON considering Toronto a U. S. city during the *Jeopardy!* challenge).

In this paper, we investigate the effects of giving noisy (i.e., partly incorrect) and incomplete feedback on an L1 text revision task. To this end, we conduct a pilot user study with German native speakers, in which we ask them to revise two texts containing a number of stylistic issues. While one group receives feedback about the potential issues, including correct, incorrect, and incomplete feedback, the second group serves as a control group, who revises the texts without any technological help.

Researching how humans deal with the outputs of language processing tools and specifically with automatically generated feedback is long overdue. Though our community achieves much progress in improving the performance on annotated gold standards, we still have limited knowledge about the usefulness of the underlying methods and techniques in a practical setting. We expect that our study makes an important contribution in this direction. From the results of our and similar studies, we envision truly *intelligent* tools that assist writers in their work, rather than forcing them to click repeatedly on "ignore this issue".

Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications, pages 42–52,
San Diego, California, June 16, 2016.

2 State of the Art

There is a vast amount of scientific literature on intelligent writing assistance and automatic text correction methods in natural language processing and especially computer-assisted language learning. To evaluate such methods, we can distinguish data-driven and user-driven approaches discussed below.

2.1 Data-driven Evaluation

The most widely accepted evaluation methodology in this area is an intrinsic setup to compare a system's output with annotated reference data. For automatically identifying language-related issues and generating corresponding corrections, the *Helping Our Own* shared tasks (Dale and Kilgarriff, 2010; 2011; Dale et al., 2012) constituted a community around this type of system evaluation, which has successfully continued at the CoNLL conferences (Ng et al., 2013; 2014) and very recently at the BEA workshop (Daudaravicius, 2015). These initiatives are completed by numerous independent evaluation studies, such as the ones by Park and Levy (2011) or Perin et al. (2012) to name just two examples.

Major challenges to this evaluation methodology are: achieving a meaningful comparison of multiple systems, properly interpreting the performance metrics, and ensuring the reliability of the reference data. Chodorow et al. (2012) discuss the comparability of grammatical error detection systems and give recommendations for best practices. Bryant and Ng (2015) pose the highly important question of what is considered high-quality error detection with regard to human performance. Obviously, the quality of the reference data directly affects the evaluation scores. Systems are penalized for detecting an actual error that remained unseen by the human annotators or suggesting a valid correction not covered by the gold standard. Inter-rater agreement measures (Artstein and Poesio, 2008) provide a useful tool to assess the reliability, but as Ng et al. (2014, p. 12) note, metrics such as the kappa coefficient do "not take into account the fact that there is often more than one valid way to correct a sentence".

We believe that data-driven evaluation of intelligent writing assistance systems is vital, but given these issues, we suggest that they should be complemented by user-driven evaluation studies.

2.2 User-driven Evaluation

The user-driven evaluation of different types of language feedback has been a major research topic in writing and language learning research, before most automatic writing assistance systems evolved. Jacobs (1989), Owston et al. (1992), and Jacobs et al. (1998) are early works in this direction discussing feedback by teachers and peers, based on different educational resources, and using different media.

More recently, the effects of giving automatically generated feedback became an important research question. Attali (2004) report a large-scale study of the *Criterion* system (Burstein et al., 2003). He automatically scores essays before and after providing automated feedback and notes an overall improvement of the writing quality when providing feedback. The study does, however, not vary the type of feedback in any way. Andersen et al. (2013) distinguish feedback at the text, sentence, and word levels and evaluate different granularities with a questionnaire. Heift and Rimrott (2008) study different ways of formulating feedback messages for spelling errors and find solution suggestions yielding improved results. In a similar line of research, Lavolette et al. (2015) compare immediate and delayed feedback and find that students more likely responded to correct feedback. Madnani et al. (2015) vary the extent of feedback messages about English preposition errors using a crowdsourcing setup. Regardless of the extent of the feedback messages, they find a learning effect in *detecting* errors over multiple writing sessions. But only participants who received correct and detailed feedback were able to *fix* more errors. They, however, note limitations of their study setup due to the unclear distribution of preposition errors and language proficiency of the crowdsourcing population. None of these works systematically varies correct, incorrect, and incomplete feedback.

The work by Nagata and Nakatani (2010) is most closely related to ours. They ask 26 language learners to write a number of essays and revise them under four experimental conditions: without any technological assistance, with recall-oriented automatic feedback, with precision-oriented automatic feedback, and with human feedback. They focus on two types of grammatical errors and find the precision-oriented feedback to maximize the learning effect of

the participants. Their work differs from the present paper in multiple ways: First, we consider a revision task of an unknown text instead of a self-written essay, which allows us to control for the number and distribution of errors over all participants. With this setup, we get in a position to compare the users' revisions systematically. Second, we consider German native speakers rather than English learners. Since most previous work is focused on English learners, we believe that addressing native speakers and other languages is an important research gap. Third, we consider stylistic rather than grammatical issues, which has not been extensively discussed before. Fourth, we are interested in the usefulness of intelligent writing assistance systems for improving the text quality rather than the learning effect of the users. Still, we are eager to compare our findings with the previous work and discuss this in section 7.

3 Goals and Hypotheses

The motivation for developing intelligent writing assistance systems is that authors get in a position to compose texts of higher quality, ideally with less effort, time, or need for manual feedback. Incomplete and noisy feedback could, however, severely hamper this goal and yield lower quality or higher effort.

To operationalize these thoughts, we simulate an intelligent writing environment that highlights stylistic issues in a text and provides brief feedback messages explaining them. We have the following four hypotheses about the usage of such a system for a text revision task:

1. If users receive correct feedback about a stylistic issue, they will more likely revise the corresponding part of a text than users, who do not receive any feedback.

2. If users receive incorrect feedback about a stylistic issue, they will more likely revise the corresponding part of a text, although this would not be necessary.

3. If users receive incomplete feedback about stylistic issues, they will more likely miss issues, for which they do not receive feedback.

4. The time required for revising the text will not significantly differ between the users of a system with and without technological assistance.

The rationale behind the first hypothesis is that the highlighted text parts direct a user's attention to the stylistic issue. We thus expect a significantly higher number of revised stylistic issues that have been highlighted to the users.

The second hypothesis follows the same motivation as the first one: The users' attention is directed to the highlighted text positions. We believe that a user will more likely revise these highlighted text parts even if this would not be necessary. This would mean that users overtrust the system, even if they are aware of potential errors in the provided feedback. We therefore expect a significantly higher number of revised text positions that do not contain a stylistic issue, but that are highlighted as such.

A different type of overtrust is that users receiving automated feedback will more likely miss issues of similar types if they are not highlighted. We thus believe that the provided feedback causes a shift of focus from the actual revision task to the processing of the highlighted text parts. In this case, we would observe a significantly lower number of unmarked revised stylistic issues if other parts of the text are highlighted and associated with feedback messages.

The fourth hypothesis considers the time required for the revision task. We expect that users receiving feedback and users not receiving feedback will take equally long and therefore no significant difference in the time to complete the task. This would mean that an intelligent writing assistant neither increases nor decreases the required revision time.

4 Experimental Design

To test our hypotheses, we conduct an empirical user study, in which we ask our participants to enhance the quality of two given texts. We employ a 2×2 mixed factorial design. That is, we divide the participants into an experimental and a control group (between-subject variable) and provide them with texts of two different text types (within-subject variable). While the control group does not receive any assistance, the experimental group receives correct, incorrect, and incomplete feedback about stylistic issues in the texts. Below, we first introduce the textual data and the types of issues we consider, before we describe the participants and the overall setup of the study.

4.1 Data

For our experiment, we require texts with a predefined set of stylistic issues. Error-annotated learner texts seem an obvious choice. However, we need texts with multiple similar issues in order to systematically compare how users deal with the different types of feedback. Therefore, we turn towards existing high-quality texts and manually introduce a number of similar stylistic issues instead of using pre-annotated (learner) texts.

We select two different text types. The first text T_1 is an excerpt of the news article "Die Zaubertafel" (Engl.: "the magical board") about the presentation of the first iPad in 2010, published by the major German newspaper *ZEIT online*.[1] Along the lines of Christensen et al. (2014), we intentionally use an old article to minimize side effects caused by prior knowledge of the participants. As the second text T_2, we use a part of the encyclopedic article "Eigentliche Pythons" (Engl.: pythons) from the German Wikipedia.[2] Both texts have about 200 words and exhibit a high text quality. At the same time, both text types also demand for a high quality. This is relevant to control for the expectations of the participants, because text types typically showing lower quality (e.g., learner essays, meeting protocols, personal notes) might not be revised to the same meticulous degree by all participants.

We manually define eleven positions $p \in [1, 11]$ within the texts as our main subjects of analysis. For eight of them, we carefully manipulate the original text to introduce a stylistic issue. The other three remain unchanged. We restrict our manipulations to three issue types:

- *inappropriate registers* (IR), such as using colloquial language in an encyclopedic article,
- *uncommon collocations* (CL), for example when using "yellow" rather than "blond" in the context of hair colors, and
- *insufficient variation* (VA) by repeatedly using the same lexical and syntactic patterns without being a rhetorical device.

We choose stylistic issues over spelling and grammar errors, since we expect automatic methods to yield even more false alarms and incorrect suggestions for them than for other issues. We discuss the manipulated texts with multiple colleagues to ensure that the introduced issues can, in principal, be recognized and fixed.

In the next step, we simulate the feedback of an intelligent writing assistance tool. That is, we highlight the words at a position p with yellow background color and we generate a message explaining the issue. Consider for example the IR issue $p = 7$:

> Wie alle Pythonartige sind sie ungiftig und machen ihrer Beute durch Umschlingen den Garaus.

The highlighted phrase "machen [...] den Garaus" is considered colloquial speech meaning to murder someone (i.e., to bump someone off). It is our manipulation of using the verb "töten" (to murder someone, without any register marking). As indicated by the example, we also allow for discontinuous highlights (i.e., a position p might refer to multiple non-adjacent words or phrases).

The corresponding feedback message for this issue is the German equivalent of:

> The phrase "*to bump so. off*" is considered colloquial speech. Check if this phrase is appropriate in the given context.

To keep the cognitive load as small as possible, we limit ourselves to brief feedback messages. The message points out that there *might* be an issue and asks the user to check if a reformulation is necessary. The feedback message does not give suggestions of how to resolve the issue, but leaves the final decision to the user. This is necessary to ensure a fair comparison with the control group with regard to our hypotheses (see section 3).

The main motivation for our work is analyzing how users deal with incomplete and noisy feedback. This is why, we do not give feedback for all issues. Rather, we distinguish between correctly highlighted parts of a text that need revision (TP), incorrectly highlighted parts of a text that do not need revision (FP), and parts of a text that need revision, but are not highlighted to a user (FN). From a tool perspective, the text positions of type TP are a correct system result (i.e., true positives), FP positions

[1] `http://www.zeit.de/2010/06/01-iPad` (published June 1, 2010; last accessed February 4, 2016)

[2] `https://de.wikipedia.org/w/?oldid=121124960` (published August 2, 2013; last accessed February 4, 2016)

p	Text	Issue	Manipulated	Highlighted	Feedback
1	T_1	VA	✓	✓	TP
2	T_1	IR	✓	✓	TP
3	T_1	IR	✓		FN
4	T_1	CL	✓	✓	TP
5	T_1	VA	✓		FN
6	T_1	VA		✓	FP
7	T_2	IR	✓	✓	TP
8	T_2	CL		✓	FP
9	T_2	IR		✓	FP
10	T_2	VA	✓	✓	TP
11	T_2	CL	✓		FN

Table 1: Text positions considered for the user study

are detected by a system but false alarms (false positives), and FN positions are errors that remain undetected by the system (false negatives).

Table 1 gives an overview of the eleven relevant text positions p. The four TPs, three FPs, and four FNs are roughly equally distributed over the two texts and the three issue types considered. The chosen stylistic issue types, their distribution and position in the text follow practical considerations induced by the underlying texts. That is to say, we aim at making minimal changes to the texts and keep their original content and organization intact, which is necessary to avoid coherence breaks and newly introduced ambiguity.

4.2 Participants

We record the revision results of 26 participants. Though finding voluntary users is notoriously difficult, we aim at reducing the bias caused by a single homogeneous user group. This is why, we ask users from three different contexts: students from different programs at our university (31 %), PhD- or Postdoc-level NLP researchers (31 %), and randomly selected volunteers with varying professional backgrounds (38 %). All participants are German native speakers, 62 % of them are male, and their age ranges from 22 to 50 with an average age of 30.2 ± 6.6. Of the 26 responses received, 15 participants revised the texts under experimental conditions and 11 were part of the control group.

4.3 Procedure

We randomly assign the participants into the two groups. Each participant receives a printout with instructions, user credentials for the writing environment, and a usability questionnaire. We first ask the participants to read the instructions, in which we explain the text revision task, the two text types, and our expectations regarding a high text quality. We ask the users to not change the meaning and organization of the texts, but to focus on stylistic issues. Both groups receive the same instruction that the writing environment *might* support the revision task and that this support is not necessarily complete or correct. No additional help or resources should be used to complete the task. To avoid any pressure for the participating students, we made clear that participation is on a voluntary basis and does not affect the grading of any course.

Having read the instructions, the participants access our online writing environment described in section 5 below. The writing environment shows the two selected texts one after another. We randomly shuffle their order to avoid effects based on the order of the two texts. Note that hereafter, we always use the original order (T_1 before T_2) for our analysis. While participants of the control group can only use common word processor functions to revise the texts, participants of the experimental group additionally see the highlighted text parts and the corresponding feedback messages according to table 1.

For performing the final step of the study, the participants save their revisions in the online system and turn towards the questionnaire printout. We record some demographic data such as age and gender as well as information about the native tongue and a self-assessment of German language skills. The main body of the questionnaire aims at studying the usability of the writing environment in order to control for side effects due to a lack of user-friendliness. In section 6, we analyze these results.

We finalize the details and the formulations of our study by conducting a pretest with a student volunteer, who is not part of the actual study participants. Based on this pretest, we clarify the formulations of the task instructions to avoid misunderstandings.

5 Writing Environment

To conduct our user study, we implement a novel web-based writing environment as a secondary contribution of this paper. The writing environment

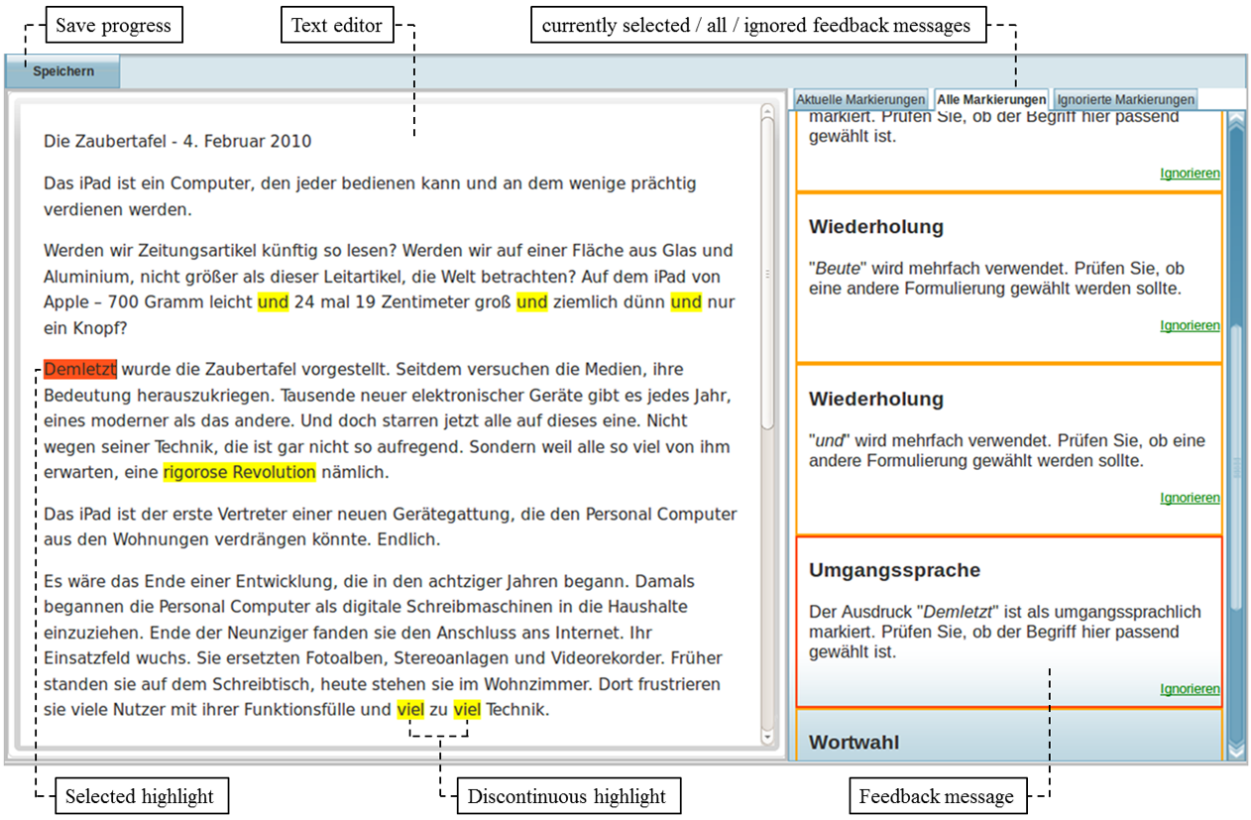

Figure 1: Screenshot of our writing environment

features common text edit operations and assists its users by displaying feedback about language-related issues. Although the highlighted text parts and the corresponding feedback messages for the stylistic issues considered in this study could also be modeled as a static webpage, we develop the writing environment with a larger goal in mind: to establish an open research platform for evaluating methods of intelligent writing assistance. The tool is available as open-source software from GitHub.[3] Figure 1 shows a screenshot of the user interface.

The writing environment divides the screen into two parts: a text editor on the left-hand side and a panel for displaying feedback on the right-hand side (about one third of the screen width). The text editor features common edit operations, such as cut/copy/paste, cursor navigation, deleting characters and selections, etc. To draw the user's attention to a certain part of the text, the editor may display words or phrases with a certain background color, similar to using a marker pen on paper. Our system can properly highlight discontinuous text parts. For the example "machen [...] den Garaus" introduced above, we can highlight the first and the second part individually without losing the link to the same feedback message. This is especially relevant for German, which is rich in separable verbs (i.e., verbs that contain a particle either as a prefix or as a separate word at the end of the sentence). Upon clicking on a highlighted text part, the background color changes to orange, indicating that this issue is currently in the focus. For discontinuous issues, we recolor all highlighted parts linked to the issue.

In the feedback panel on the right-hand side of the screen, the user can choose to view a list of all feedback messages (tab "Alle Markierungen") or only the currently selected ones (tab "Aktuelle Markierungen", default setting). Note that in a real usage scenario, there could be multiple overlapping issues, which is why the current selection may include more than a single feedback message. Clicking on a feedback message has the same effect as

[3] https://github.com/UKPLab/naacl-bea2016-writing-study

clicking on a highlighted text part – the system will mark both the text part and the feedback message in orange. When editing a highlighted text part, the yellow background color will disappear, similar to the spell-checking functionality of common word processors. Users may optionally ignore an issue without editing it. This clears the background color and moves the feedback message to a separate tab "Ignorierte Markierungen", from where it can also be reactivated in case it was ignored by accident or saved for later.

A key feature of our writing environment is that all user–system interactions are recorded and sent to a server instance, where we can analyze and store them. Specifically, we can log the keystrokes, the cursor navigation, and the interaction with the highlighted text parts and the feedback messages. Since each recorded interaction has a timestamp, we get in a position to determine the time to complete a certain writing task or phase. The recorded user–system interaction data for the revision task described above is the data basis for checking our hypotheses.

6 System Usability

To rule out that the measured effects are influenced by a bad design of the writing environment, we ask our participants in the experimental group to rate the system usability.

The *System Usability Scale* (SUS) introduced by Brooke (1996) is among the most widely used measures. The SUS score is based on the user ratings for ten questions using a five point Likert scale each. For a given user u, the score is defined as

$$SUS(u) = 2.5 \Big(\sum_{i \in \{1,3,5,7,9\}} u_i \; + \sum_{i \in \{2,4,6,8,10\}} 4 - u_i \Big)$$

where $u_i = 0 \ldots 4$ is the user's rating for question i. Typically, the individual SUS scores are averaged over all users.

For our study, we use the German SUS translation by Lohmann and Schäfer (2013). One user skipped question 6 ("I thought there was too much inconsistency in this system") and another user skipped the questions 6 and 2 ("I found the system unnecessarily complex") for which we assume the neutral score 2.

Our system achieves an average SUS score of 76.3. While 100 is the maximum score, 68 is considered the threshold between poor and acceptable usability. For scores between 71.4 and 85.5, Bangor et al. (2009) find the highest correlation with the adjective "good", which is why we conclude that the observations made during our study are not affected by a poor system usability.

7 Results

For checking our four hypotheses, we identify which participant revised the texts at each of the eleven text positions p. All native speaker revisions yielded acceptable texts, which is why we consider a revision at p on a binary scale. This provides us with a total of $11 \cdot 26 = 286$ data points for our analysis; 165 for the experimental group and 121 for the control group. On average, participants of the experimental group revised $\overline{x}_{\mathrm{EG}} = 5.86$ (standard error $SE = 0.53$, min: 2, max: 10) text positions and participants of the control group $\overline{x}_{\mathrm{CG}} = 3.18$ ($SE = 0.74$, min: 0, max: 8). Figure 2 (a) shows a notched boxplot of the total number of revised text positions. In addition to that, we consider the 26 times (in seconds) to complete the task. To test the hypotheses, we use an unpaired two sample Student's t-test and a significance level of $\alpha = 0.05$ (i.e., $P \leq 0.05$).

7.1 Hypothesis 1: Correct feedback

Our first hypothesis is that participants of the experimental group will more likely revise text parts that are highlighted compared to the control group not receiving any highlights. The corresponding null hypothesis is that $\mu_{\mathrm{EG(TP)}} = \mu_{\mathrm{CG(TP)}}$, where $\mu_{\mathrm{EG(TP)}}$ denotes the expected value of the number of changes at TP positions made by the experimental group and $\mu_{\mathrm{CG(TP)}}$ the corresponding expected value for the control group.

The mean number of revisions of TP positions $\{1, 2, 4, 7, 10\}$ is $\overline{x}_{\mathrm{EG(TP)}} = 4.13$ ($SE = 0.23$) for the experimental group and $\overline{x}_{\mathrm{CG(TP)}} = 1.63$ ($SE = 0.51$) for the control group. All participants in the experimental group revised at least 2 positions, while there are 4 participants of the control group who did not revise a single TP position. Conversely, there are participants of both groups, who revised all 5 TP positions. Figure 2 (b) shows a boxplot indicating a higher number of revisions in the experimental group than in the control group, whereas the control group shows a higher variance.

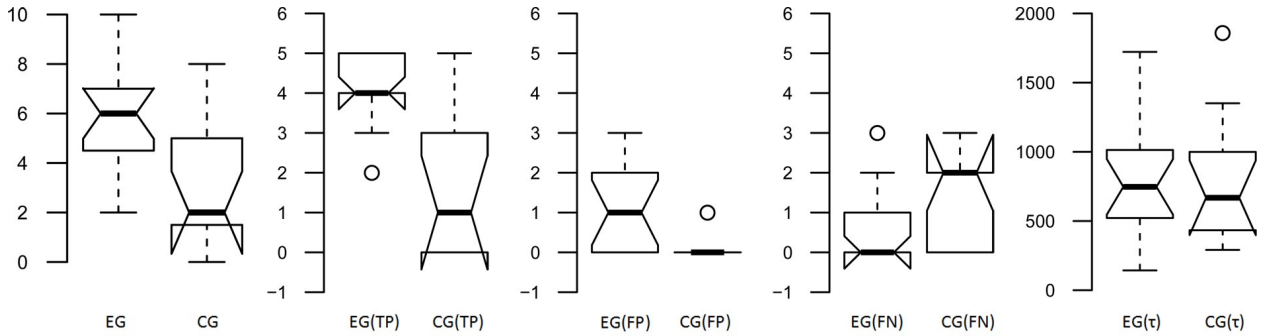

Figure 2: Notched boxplots ($\pm 1.57 \frac{IQR}{\sqrt{n}}$) comparing all revised positions (a), the revisions at TP positions (b), at FP positions (c), at FN positions (d), and the time to complete the task (e) of the experimental group (EG) and the control group (CG)

The test statistic computes to $t_1 = 4.85$, which is higher than the critical value 2.06 ($P < 0.0001$). We can therefore clearly reject the null hypothesis at the 5 % level in favor of the alternative that highlighting stylistic issues helps the participants with increasing the text quality.

7.2 Hypothesis 2: Noisy feedback

Our second hypothesis is that the participants will more likely revise text positions that are mislabeled as stylistic issues. In other words, we expect a significant difference in the number of times the experimental group revises the FP positions $\{6, 8, 9\}$ compared to the control group. The corresponding null hypothesis is that $\mu_{\mathrm{EG(FP)}} = \mu_{\mathrm{CG(FP)}}$ where $\mu_{\mathrm{EG(FP)}}$ denotes the expected value of the number of changes at FP positions made by the experimental group and $\mu_{\mathrm{CG(FP)}}$ the corresponding expected value for the control group.

The mean number of revisions of FP positions is $\overline{x}_{\mathrm{EG(FP)}} = 1$ ($SE = 0.25$) for the experimental group and $\overline{x}_{\mathrm{CG(FP)}} = 0.18$ ($SE = 0.12$ for the control group. There are participants in both groups who did not revise any FP position. In the control group, only two participants revised a single FP position at all. In the experimental group, four participants revised a single, another four revised two, and one participant even revised all three FP positions, which corroborates our hypothesis. Figure 2 (c) shows the boxplot for FP positions.

We compute the test statistic $t_2 = 2.55$, which is higher than the critical value 2.06 ($P = 0.017$). We can therefore reject the null hypothesis at the 5 % level and conclude that highlighting false alarms causes writers to unnecessarily edit their manuscript.

While the results reported so far might be considered obvious, we note that the group difference is less clear than expected and much smaller than the one for the first hypothesis. Since six participants of the experimental group were able to recognize and ignore all false alarms, we suggest that intelligent methods should take the user interaction into account and control for the internal thresholds controlling the precision–recall trade-off. That is to say, users accepting all or most suggestions of a system, including those with a low confidence, should receive a higher precision, whereas users carefully picking out what to revise might be interested in a higher recall. This goes beyond Nagata and Nakatani's (2010) precision-focused suggestion.

7.3 Hypothesis 3: Incomplete feedback

Our third hypothesis is that the participants whose texts contain highlighted parts will rather not recognize stylistic issues of a similar type if they are not highlighted as well. In other words, we expect a significant difference in the number of times the participants of either group revise the FN positions $\{3, 5, 11\}$. The corresponding null hypothesis is that $\mu_{\mathrm{EG(FN)}} = \mu_{\mathrm{CG(FN)}}$ where $\mu_{\mathrm{EG(FN)}}$ denotes the expected value of the number of changes at FN positions made by the experimental group and $\mu_{\mathrm{CG(FN)}}$ the corresponding expected value for the control group.

The mean number of revisions of FN positions is $\overline{x}_{\mathrm{EG(FN)}} = 0.73$ ($SE = 0.28$) for the experimental group and $\overline{x}_{\mathrm{CG(FN)}} = 1.36$ ($SE = 0.36$) for the control group. Both groups contain participants,

who revised either all three FN positions or none of them. Figure 2 (d) shows the corresponding boxplot.

Although the notches of the boxplot do not overlap (indicating a statistical difference), the test statistic is $t_3 = -1.39$, whose absolute value is clearly lower than the critical value 2.06 ($P = 0.17$). We therefore cannot reject the null hypothesis and thus do not find a significant difference between the two groups. This means that although we note a tendency for users seeing highlighted text parts to overlook unmarked issues of the same type, we do not find a significant difference.

While future studies with a larger number of participants may find a significant difference (mind that we cannot reject the alternative hypothesis based on our results), we note that false positives seem to be a more severe problem when giving automatic writing feedback than false negatives. For the design of an intelligent writing assistance systems, we therefore agree to Nagata and Nakatani (2010) in that we should particularly focus on precision (i.e., avoid false alarms) before aiming at an optimized recall.

7.4 Hypothesis 4: Task completion time

Our final hypothesis states that participants of the experimental group do not take significantly longer to complete the task than participants of the control group. We therefore expect that $\mu_{\mathrm{EG}(\tau)} = \mu_{\mathrm{CG}(\tau)}$ where $\mu_{\mathrm{EG}(\tau)}$ is the expected value of the task completion time of the experimental group and $\mu_{\mathrm{CG}(\tau)}$ correspondingly of the control group.

The task completion times range from 2 min, 23 sec to 30 min, 59 sec. The majority of participants require between 7 and 16 min with a mean of $\overline{x}_{\mathrm{EG}(\tau)} = 13$ min, 3 sec ($SE = 104$ sec) for the experimental group and $\overline{x}_{\mathrm{CG}(\tau)} = 13$ min, 27 sec ($SE = 144$ sec) for the control group, which is surprisingly similar. Figure 2 (e) shows again a boxplot.

The test statistic is $t_4 = -0.14$. The absolute value is clearly lower than the critical value 2.06 ($P = 0.89$). We therefore cannot reject the null hypothesis and thus do not find a significant difference between the two groups.

If there is no significant difference in the time that is required to revise a text with and without automatic feedback, this is good news for building intelligent writing assistance tools, as they do not cause additional expenditure of time for the writers.

8 Conclusion and Future Work

We conducted an empirical user study to analyze the effects of assisting writers with incomplete and noisy feedback when revising a given text. To this end, we systematically introduced stylistic issues in two texts and asked voluntary participants to enhance the quality of the texts. An experimental group received technological assistance by means of highlighted issues and corresponding feedback messages. We distinguished between highlighted text parts that needed revision (TP), highlighted text parts that did not require revision (FP), and texts parts that required revision without being highlighted (FN). With this setup, we simulated the error types of an actual intelligent writing assistance system. We compared the performance of the experimental group to a control group, who did not receive any technological aids.

Our analysis revealed that highlighting stylistic issues helped the participants to improve the quality of a text. If a text part was highlighted, the participants more likely revised it, even if the given text was already correct. In contrast, we found no significant difference for issues that remained undetected by a system (i.e., incomplete feedback). We concluded that the precision of a system plays a more important role than its recall, as participants tend to overtrust the system output, even though we made clear that the given feedback is not necessarily correct.

As a secondary contribution, we describe a novel writing environment, which we used for our study. We found a good system usability score for the tool and did not find a significant difference in the time to complete the text revision task indicating that neither the tool nor the feedback hinders the task.

We consider the user-driven evaluation of intelligent writing assistance and automatic text correction systems as highly important for assessing their usefulness. Follow-up studies should vary the frequency, order, and distribution of the issues and experiment with different ways of giving feedback. Based on our study results, we consider adaptive and interactive methods highly promising for designing and evaluating intelligent writing assistance tools. Besides writing assistance, future advances are also relevant for automatic essay scoring tools, which could allow for a more fine-grained analysis.

Acknowledgments

This work has been supported by the DFG-funded research training group "Adaptive Preparation of Information form Heterogeneous Sources" (AIPHES, GRK 1994/1) and by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant № I/82806. We would like to thank the anonymous reviewers for their helpful comments.

References

Øistein E. Andersen, Helen Yannakoudakis, Fiona Barker, and Tim Parish. 2013. Developing and testing a self-assessment and tutoring system. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 32–41, Atlanta, GA, USA.

Ron Artstein and Massimo Poesio. 2008. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596.

Yigal Attali. 2004. Exploring the Feedback and Revision Features of Criterion. Paper presented at the National Council on Measurement in Education (NCME), April 12–16, 2004, San Diego, CA, USA.

Aaron Bangor, Philip Kortum, and James Miller. 2009. Determining What Individual SUS Scores Mean: Adding an Adjective Rating Scale. *Journal of Usability Studies*, 4(3):114–123.

John Brooke. 1996. SUS: A 'Quick and Dirty' Usability Scale. In Patrick W. Jordan, Bruce Thomas, Bernard A. Weerdmeester, and Ian L. McClelland, editors, *Usability evaluation in industry*, chapter 21, pages 189–194. London//Bristol, PA: Taylor & Francis.

Christopher Bryant and Hwee Tou Ng. 2015. How Far are We from Fully Automatic High Quality Grammatical Error Correction? In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing: Long Papers*, pages 697–707, Beijing, China.

Jill Burstein, Martin Chodorow, and Claudia Leacock. 2003. Criterion Online Essay Evaluation: An Application for Automated Evaluation of Student Essays. In *Proceedings of the Fifteenth Annual Conference on Innovative Applications of Artificial Intelligence*, pages 3–10, Acapulco, Mexico.

Martin Chodorow, Markus Dickinson, Ross Israel, and Joel Tetreault. 2012. Problems in Evaluating Grammatical Error Detection Systems. In *Proceedings of the 24th International Conference on Computational Linguistics*, volume 2, pages 611–628, Mumbai, India.

Janara Christensen, Stephen Soderland, Gagan Bansal, and Mausam. 2014. Hierarchical Summarization: Scaling Up Multi-Document Summarization. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 902–912, Baltimore, MD, USA.

Robert Dale and Adam Kilgarriff. 2010. Helping Our Own: Text Massaging for Computational Linguistics as a New Shared Task. In *Proceedings of the 6th International Natural Language Generation Conference*, pages 261–265, Dublin, Ireland.

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, Nancy, France.

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, Montréal, QC, Canada.

Vidas Daudaravicius. 2015. Automated Evaluation of Scientific Writing: AESW Shared Task Proposal. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 56–63, Denver, CO, USA.

Trude Heift and Anne Rimrott. 2008. Learner responses to corrective feedback for spelling errors in CALL. *System*, 36(2):196–213.

George M. Jacobs, Andy Curtis, George Braine, and Su-Yueh Huang. 1998. Feedback on student writing: taking the middle path. *Journal of Second Language Writing*, 7(3):307–317.

George M. Jacobs. 1989. Dictionaries Can Help Writing – If Students Know How To Use Them. ERIC Document Reproduction Service ED 316 025, Department of Educational Psychology, University of Hawaii.

Elizabeth Lavolette, Charlene Polio, and Jimin Kahng. 2015. The Accuracy of Computer-Assisted Feedback and Students' Responses to It. *Language Learning & Technology*, 19(2):50–68.

Kris Lohmann and Jörg Schäfer. 2013. System Usability Scale (SUS) – An Improved German Translation of the Questionnaire. Online: http://minds.coremedia.com/2013/09/18/sus-scale-an-improved-german-translation-questionnaire, September, 18, 2013. (accessed: February 4, 2016).

Nitin Madnani, Martin Chodorow, Aoife Cahill, Melissa Lopez, Yoko Futagi, and Yigal Attali. 2015. Preliminary Experiments on Crowdsourced Evaluation of Feedback Granularity. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 162–171, Denver, CO, USA.

Ryo Nagata and Kazuhide Nakatani. 2010. Evaluating performance of grammatical error detection to maximize learning effect. In *Proceedings of the 23rd International Conference on Computational Linguistics: Poster Volume*, pages 894–900, Beijing, China.

Hwee Tou Ng, Siew Mei Wu, Yuanbin Wu, Christian Hadiwinoto, and Joel Tetreault. 2013. The CoNLL-2013 Shared Task on Grammatical Error Correction. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–12, Sofia, Bulgaria.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. The CoNLL-2014 Shared Task on Grammatical Error Correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14, Baltimore, MD, USA.

Ronald D. Owston, Sharon Murphy, and Herbert H. Wideman. 1992. The Effects of Word Processing on Students' Writing Quality and Revision Strategies. *Research in the Teaching of English*, 26(3):249–276.

Y. Albert Park and Roger Levy. 2011. Automated Whole Sentence Grammar Correction Using a Noisy Channel Model. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 934–944, Portland, OR, USA.

Fabrizio Perin, Lukas Renggli, and Jorge Ressia. 2012. Linguistic style checking with program checking tools. *Computer Languages, Systems & Structures*, 38(1):61–72.