Computer-assisted stylistic revision with incomplete and noisy feedback

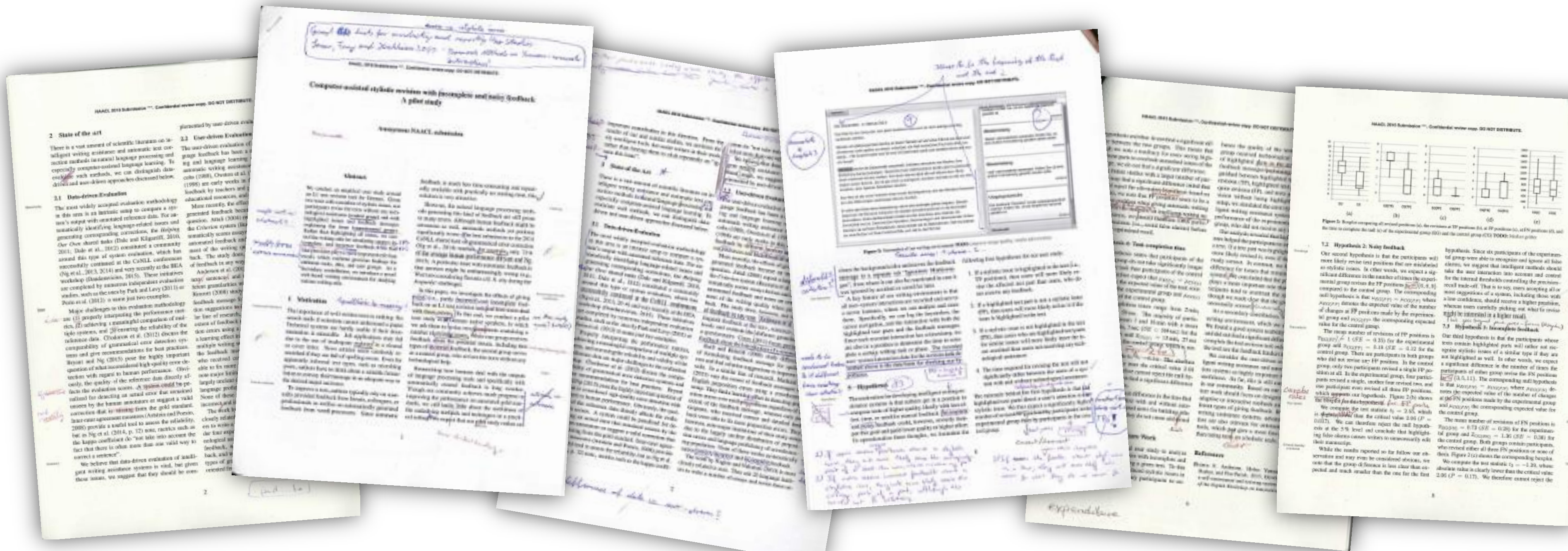
A pilot study

Christian M. Meyer and **Johann Frerik Koch**

The 11th Workshop on Innovative Use of NLP
for Building Educational Applications (BEA)
June 16, 2016. San Diego, CA, USA.

Vision

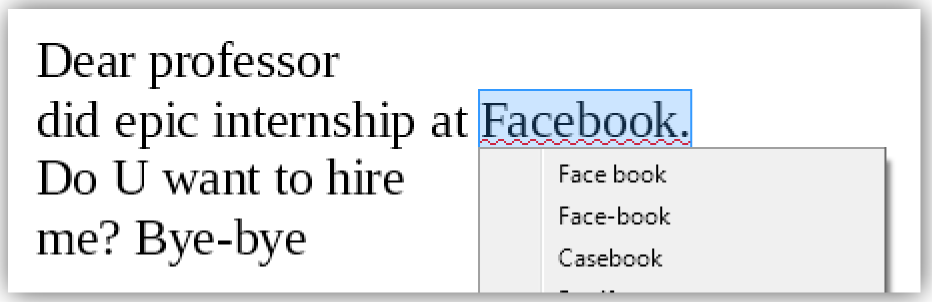
- Bryant & Ng (2015): best grammar correction software achieves only 73% of human performance
- **Our vision:** research new <u>useful</u> approaches to intelligent writing assistance with a focus on German native speakers



Goal of this work

There won’t be perfect systems! ☹

How do users deal with incorrect and incomplete feedback?



Dear professor
did epic internship at Facebook.
Do U want to hire
me? Bye-bye

- **Pilot user study**
- German L1 text revision task
- focus on stylistic issues

Previous Work

Data-driven evaluation

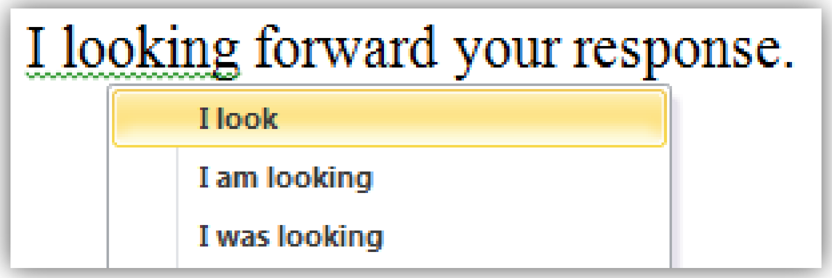
- Shared tasks: HOO, CoNLL, BEA,…
- meaningful system comparison?
- interpretation of evaluation metrics?
- reliability of the reference data?

User-driven evaluation

- (Manual) feedback by teachers and peers
- Variation of feedback granularity, extent & formulation, time
- Nagata & Nakatani (2010): *“precision-oriented error detection is better than recall-oriented”*

Hypotheses

H1 If users receive **correct feedback**, they will more likely revise the corresponding section

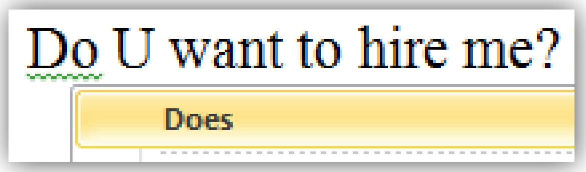


I looking forward your response.

Hypotheses

H1 If users receive **correct feedback**, they will more likely revise the corresponding section

H2 If users receive **incorrect feedback**, they will more likely revise the corresponding section
– although it would not be necessary



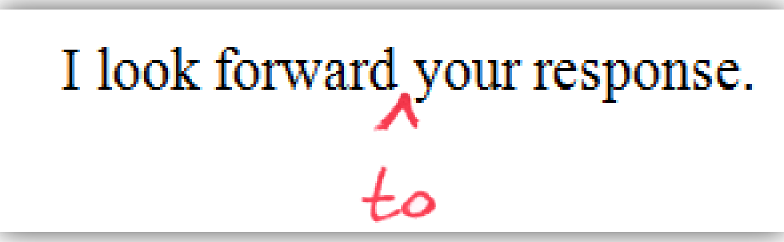
Do U want to hire me?

Hypotheses

H1 If users receive **correct feedback**, they will more likely revise the corresponding section

H2 If users receive **incorrect feedback**, they will more likely revise the corresponding section
– although it would not be necessary

H3 If users receive **incomplete feedback**, they will more likely miss issues not highlighted to them



I look forward ^ your response.

to

Hypotheses

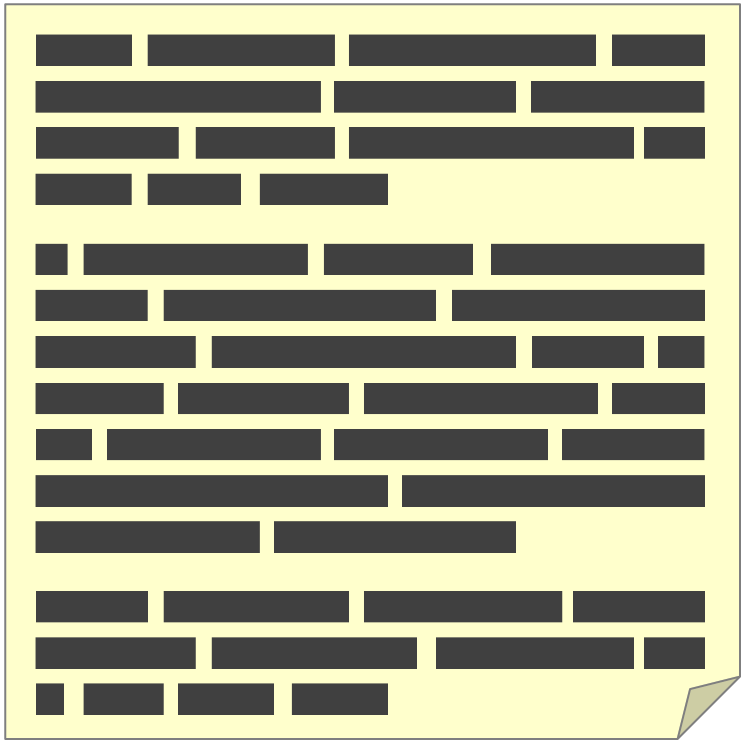
H1 If users receive **correct feedback**, they will more likely revise the corresponding section

H2 If users receive **incorrect feedback**, they will more likely revise the corresponding section
– although it would not be necessary

H3 If users receive **incomplete feedback**, they will more likely miss issues not highlighted to them

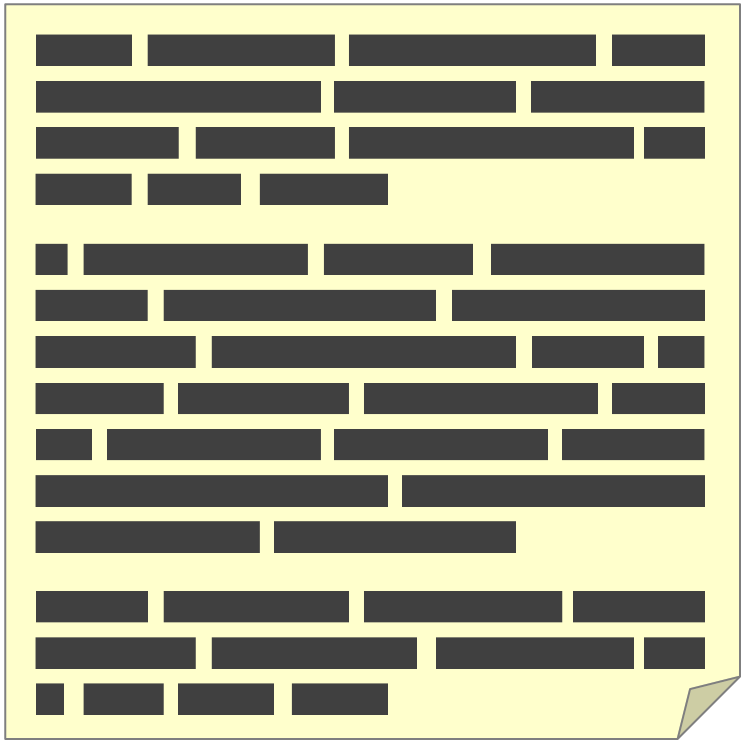
H4 Providing automatic feedback does not affect the required **time to complete the task**

Experimental Setup: Data



***T*₁ News item**

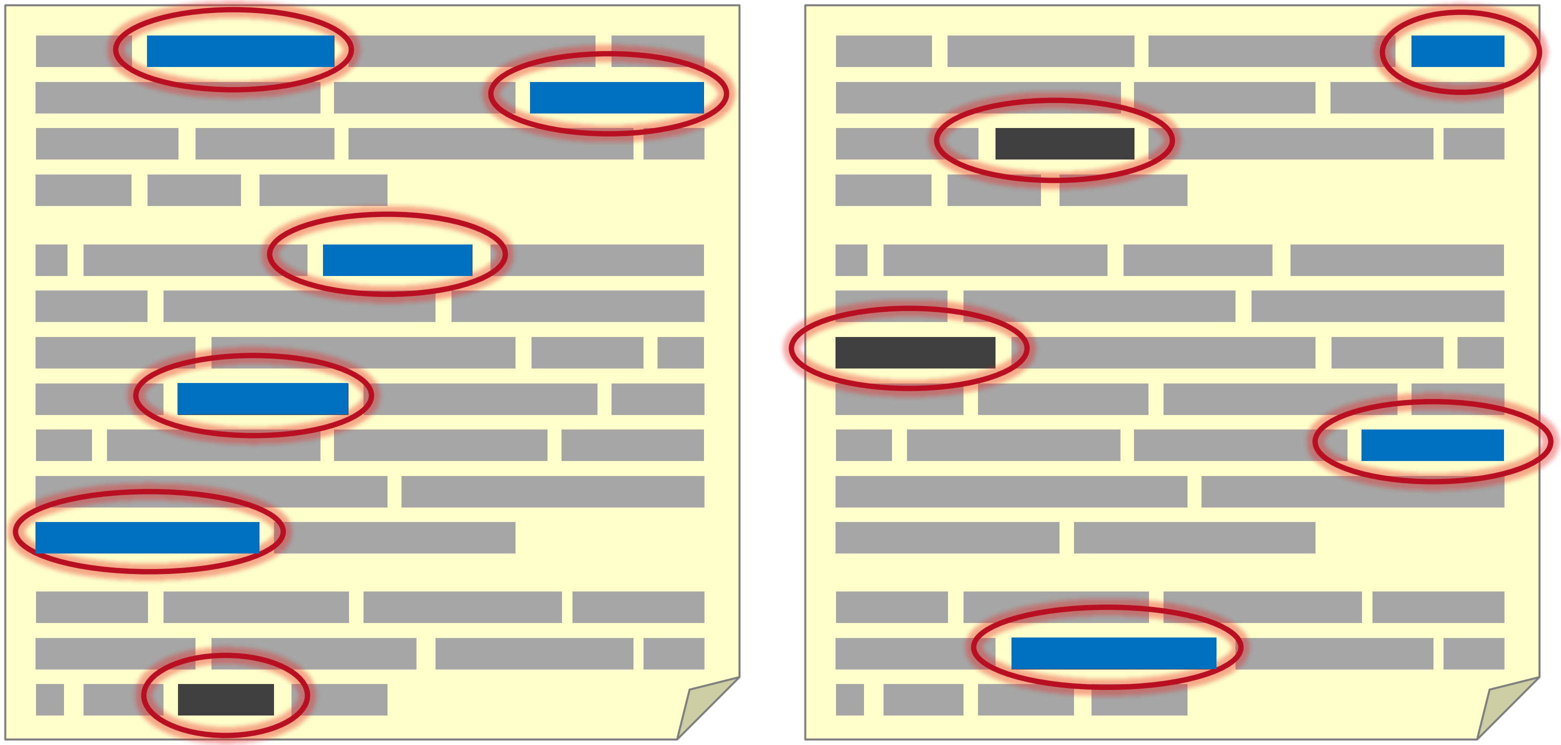
206 words



***T*₂ Wikipedia article**

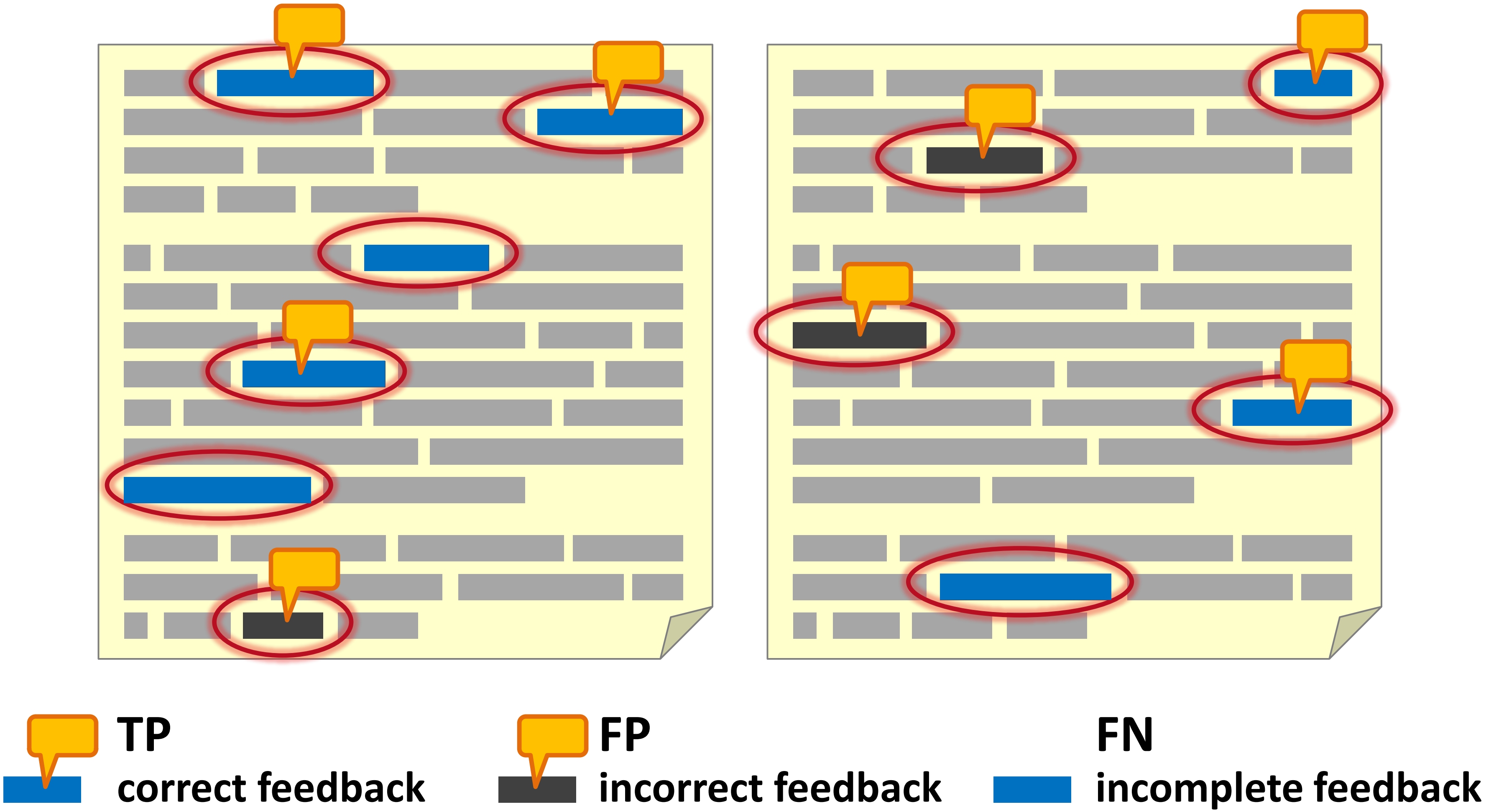
183 words

Experimental Setup: Data



11 text positions   **8 introduced issues**

Experimental Setup: Data

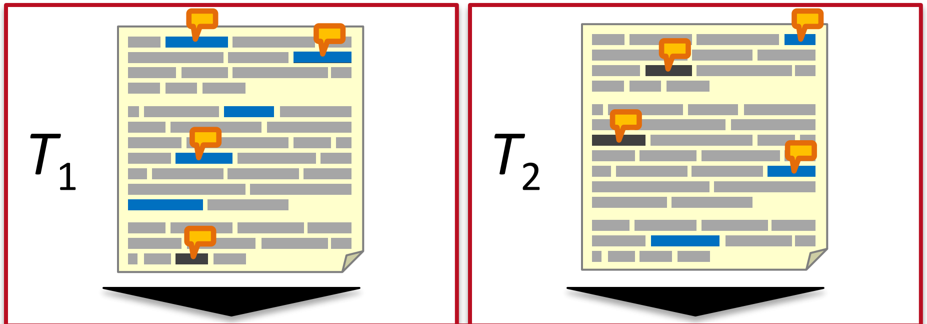


TP
correct feedback

FP
incorrect feedback

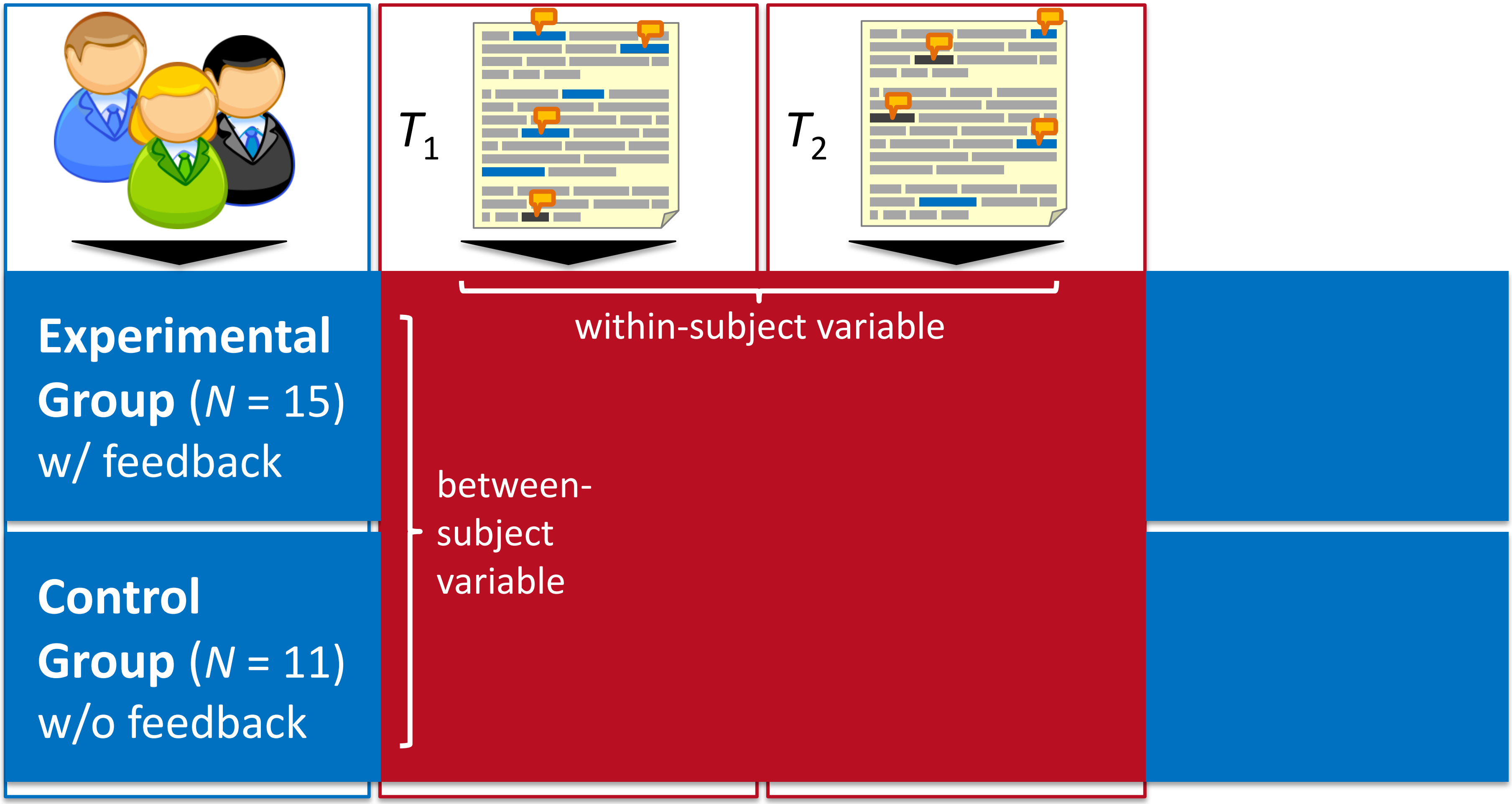
FN
incomplete feedback

Experimental Setup: Data



Issue	TP	FP	FN
Register	2	1	1
Collocation	1	1	1
Variation	2	1	1

Experimental Setup: Population

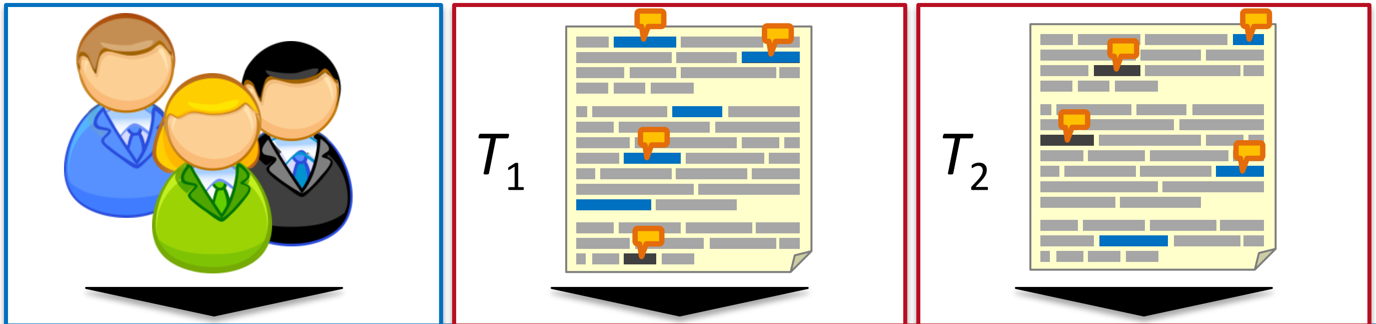


Experimental Group (N = 15) w/ feedback

Control Group (N = 11) w/o feedback

https://pixabay.com/de/person-benutzer-avatar-mann-anzug-311134/ (Public Domain, CC0), https://pixabay.com/de/gesichtslosen-asiatische-schwarz-309883/ (Public Domain, CC0), https://pixabay.com/de/frau-anzug-gesch%C3%A4ft-gr%C3%BCn-m%C3%A4dchen-303109/ (Public Domain, CC0)

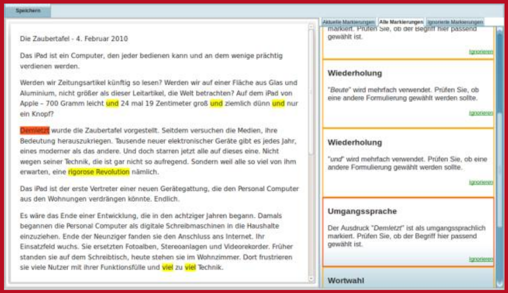
Experimental Setup: Tool



Experimental Group (*N* = 15) w/ feedback

Control Group (*N* = 11) w/o feedback

User Study



Die Zaubertafel - 4. Februar 2010

Das iPad ist ein Computer, den jeder bedienen kann und an dem wenige prächtig verdienen werden.

Werden wir Zeitungsartikel künftig so lesen? Werden wir auf einer Fläche aus Glas und Aluminium, nicht größer als dieser Leitartikel, die Welt betrachten? Auf dem iPad von Apple – 700 Gramm leicht und 24 mal 19 Zentimeter groß und ziemlich dünn und nur ein Knopf?

Demletzt wurde die Zaubertafel vorgestellt. Seitdem versuchen die Medien, ihre Bedeutung herauszukriegen. Tausende neuer elektronischer Geräte gibt es jedes Jahr, eines moderner als das andere. Und doch starren jetzt alle auf dieses eine. Nicht wegen seiner Technik, die ist gar nicht so aufregend. Sondern weil alle so viel von ihm erwarten, eine rigorose Revolution nämlich.

Das iPad ist der erste Vertreter einer neuen Gerätegattung, die den Personal Computer aus den Wohnungen verdrängen könnte. Endlich.

Es wäre das Ende einer Entwicklung, die in den achtziger Jahren begann. Damals begannen die Personal Computer als digitale Schreibmaschinen in die Haushalte einzuziehen. Ende der Neunziger fanden sie den Anschluss ans Internet. Ihr Einsatzfeld wuchs. Sie ersetzten Fotoalben, Stereoanlagen und Videorekorder. Früher standen sie auf dem Schreibtisch, heute stehen sie im Wohnzimmer. Dort frustrieren sie viele Nutzer mit ihrer Funktionsfülle und viel zu viel Technik.

markiert. Prüfen Sie, ob der Begriff hier passend gewählt ist.

Wiederholung

"*Beute*" wird mehrfach verwendet. Prüfen Sie, ob eine andere Formulierung gewählt werden sollte.

Wiederholung

"*und*" wird mehrfach verwendet. Prüfen Sie, ob eine andere Formulierung gewählt werden sollte.

Umgangssprache

Der Ausdruck "*Demletzt*" ist als umgangssprachlich markiert. Prüfen Sie, ob der Begriff hier passend gewählt ist.

Wortwahl

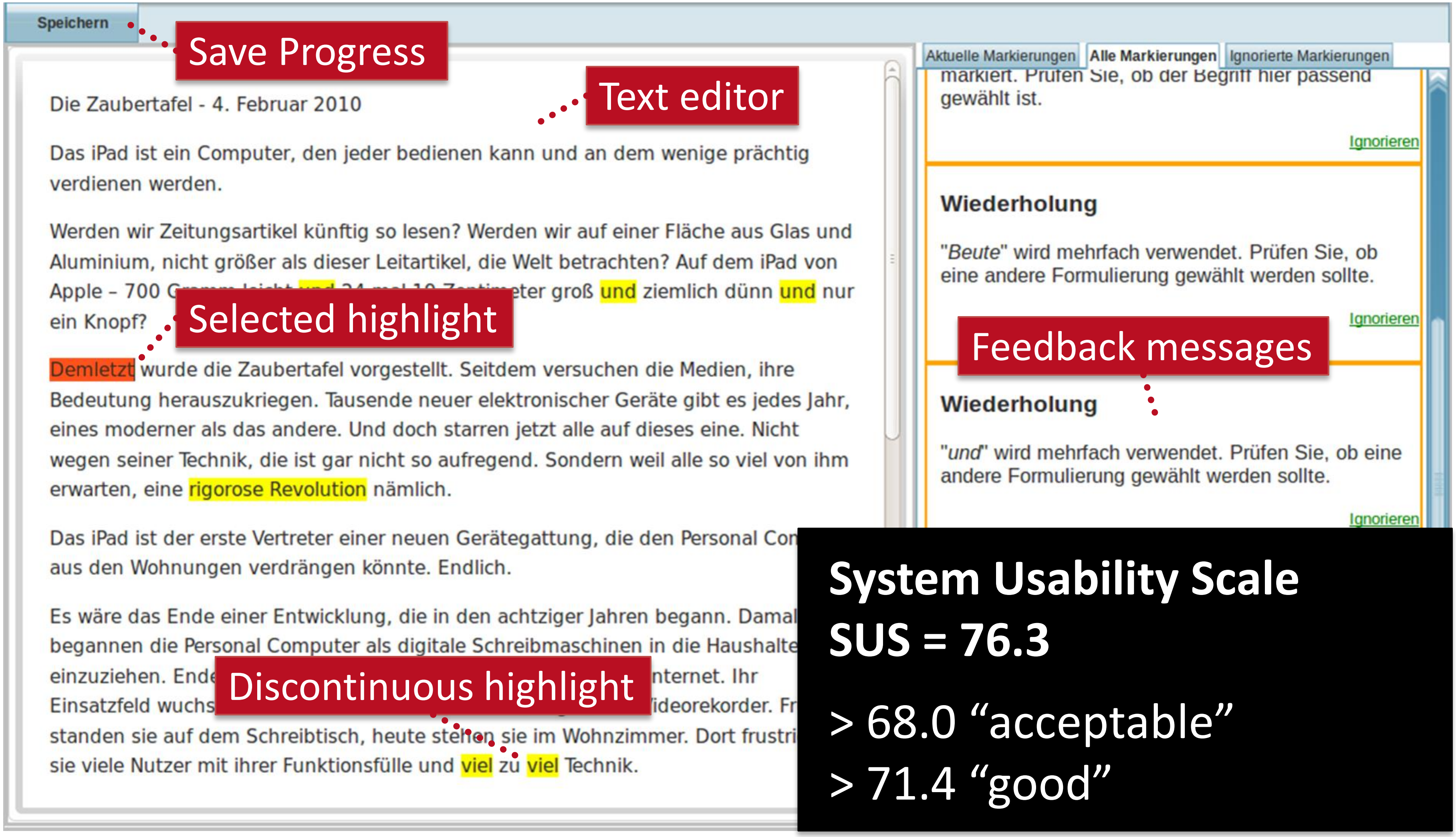
New tool: InViEdit

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

https://pixabay.com/de/person-benutzer-avatar-mann-anzug-311134/ (Public Domain, CC0), https://pixabay.com/de/gesichtslosen-asiatische-schwarz-309883/ (Public Domain, CC0), https://pixabay.com/de/frau-anzug-gesch%C3%A4ft-gr%C3%BCn-m%C3%A4dchen-303109/ (Public Domain, CC0)

Writing Assistance Software

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



Die Zaubertafel - 4. Februar 2010

Das iPad ist ein Computer, den jeder bedienen kann und an dem wenige prächtig verdienen werden.

Werden wir Zeitungsartikel künftig so lesen? Werden wir auf einer Fläche aus Glas und Aluminium, nicht größer als dieser Leitartikel, die Welt betrachten? Auf dem iPad von Apple – 700 Gramm leicht und 24 mal 10 Zentimeter groß und ziemlich dünn und nur ein Knopf?

Demletzt wurde die Zaubertafel vorgestellt. Seitdem versuchen die Medien, ihre Bedeutung herauszukriegen. Tausende neuer elektronischer Geräte gibt es jedes Jahr, eines moderner als das andere. Und doch starren jetzt alle auf dieses eine. Nicht wegen seiner Technik, die ist gar nicht so aufregend. Sondern weil alle so viel von ihm erwarten, eine rigorose Revolution nämlich.

Das iPad ist der erste Vertreter einer neuen Gerätegattung, die den Personal Con aus den Wohnungen verdrängen könnte. Endlich.

Es wäre das Ende einer Entwicklung, die in den achtziger Jahren begann. Damal begannen die Personal Computer als digitale Schreibmaschinen in die Haushalte einzuziehen. Ende nternet. Ihr Einsatzfeld wuchs ideorekorder. Fr standen sie auf dem Schreibtisch, heute stehen sie im Wohnzimmer. Dort frustri sie viele Nutzer mit ihrer Funktionsfülle und viel zu viel Technik.

markiert. Prüfen Sie, ob der Begriff hier passend gewählt ist.

Ignorieren

Wiederholung

"*Beute*" wird mehrfach verwendet. Prüfen Sie, ob eine andere Formulierung gewählt werden sollte.

Ignorieren

Wiederholung

"*und*" wird mehrfach verwendet. Prüfen Sie, ob eine andere Formulierung gewählt werden sollte.

Ignorieren

Save Progress

Text editor

Selected highlight

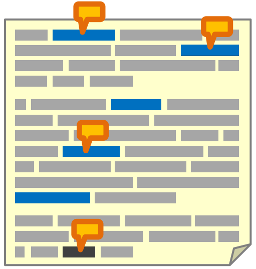
Feedback messages

Discontinuous highlight

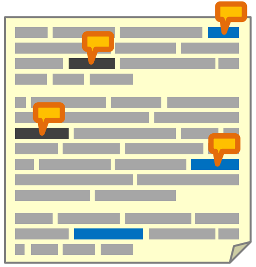
System Usability Scale
SUS = 76.3

> 68.0 “acceptable”
> 71.4 “good”

Experimental Setup: Analysis



T_1

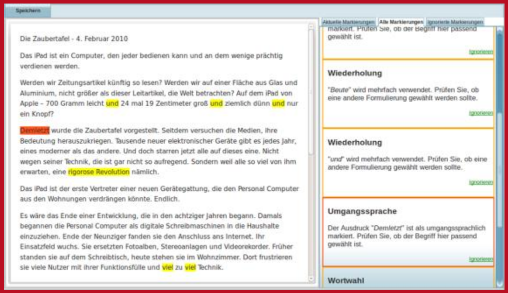


T_2

Experimental Group (N = 15) w/ feedback

Control Group (N = 11) w/o feedback

User Study



Die Zaubertafel - 4. Februar 2010

Das iPad ist ein Computer, den jeder bedienen kann und an dem wenige prächtig verdienen werden.

Werden wir Zeitungsartikel künftig so lesen? Werden wir auf einer Fläche aus Glas und Aluminium, nicht größer als dieser Leitartikel, die Welt betrachten? Auf dem iPad von Apple – 700 Gramm leicht und 24 mal 19 Zentimeter groß und ziemlich dünn und nur ein Knopf?

Demletzt wurde die Zaubertafel vorgestellt. Seitdem versuchen die Medien, ihre Bedeutung herauszukriegen. Tausende neuer elektronischer Geräte gibt es jedes Jahr, eines moderner als das andere. Und doch starren jetzt alle auf dieses eine. Nicht wegen seiner Technik, die ist gar nicht so aufregend. Sondern weil alle so viel von ihm erwarten, eine rigorose Revolution nämlich.

Das iPad ist der erste Vertreter einer neuen Gerätegattung, die den Personal Computer aus den Wohnungen verdrängen könnte. Endlich.

Es wäre das Ende einer Entwicklung, die in den achtziger Jahren begann. Damals begannen die Personal Computer als digitale Schreibmaschinen in die Haushalte einzuziehen. Ende der Neunziger fanden sie den Anschluss ans Internet. Ihr Einsatzfeld wuchs. Sie ersetzten Fotoalben, Stereoanlagen und Videorekorder. Früher standen sie auf dem Schreibtisch, heute stehen sie im Wohnzimmer. Dort frustrieren sie viele Nutzer mit ihrer Funktionsfülle und viel zu viel Technik.

markiert. Prüfen Sie, ob der Begriff hier passend gewählt ist.

Wiederholung

"*Beute*" wird mehrfach verwendet. Prüfen Sie, ob eine andere Formulierung gewählt werden sollte.

Wiederholung

"*und*" wird mehrfach verwendet. Prüfen Sie, ob eine andere Formulierung gewählt werden sollte.

Umgangssprache

Der Ausdruck "*Demletzt*" ist als umgangssprachlich markiert. Prüfen Sie, ob der Begriff hier passend gewählt ist.

Wortwahl

New tool: InViEdit

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

revised vs. not revised positions

revised vs. not revised positions

H1–H4

https://pixabay.com/de/person-benutzer-avatar-mann-anzug-311134/ (Public Domain, CC0), https://pixabay.com/de/gesichtslosen-asiatische-schwarz-309883/
(Public Domain, CC0), https://pixabay.com/de/frau-anzug-gesch%C3%A4ft-gr%C3%BCn-m%C3%A4dchen-303109/ (Public Domain, CC0)

Data Analysis

11 positions (TP/FP/FN)
× 26 participants
= 286 data points

Data point *x* = (revised vs. not revised)

	min(*x*)	$\bar{x}$	*SE*	*max(x)*
EG	2	5.86	0.53	10
CG	0	3.18	0.74	8

Unpaired two sample Student’s *t* test
with significance level $\alpha = 0.05$ ($P \leq 0.05$)

H1: Correct Feedback helps

Expectation: $\mu_{EG(TP)} \neq \mu_{CG(TP)}$

Arithmetic mean:

$\bar{x}_{EG(TP)} = 4.13$ (SE = 0.23)

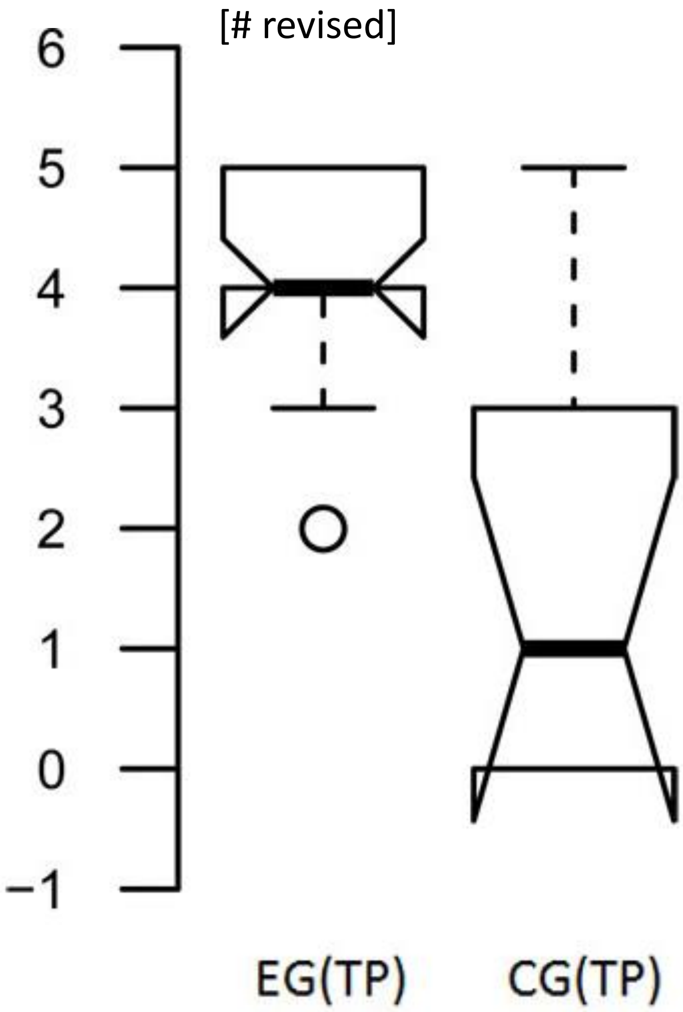
$\bar{x}_{CG(TP)} = 1.63$ (SE = 0.51)

Test statistic:

$t_{H1} = 4.85$

$|t_{H1}| > 2.06$ ($P < 0.0001$)

- reject null hypothesis at 5% level
- significant difference b/w groups ✓



H2: Incorrect Feedback causes unnecessary revisions

Expectation: $\mu_{EG(FP)} \neq \mu_{CG(FP)}$

Arithmetic mean:

$\bar{x}_{EG(FP)} = 1$ (SE = 0.25)

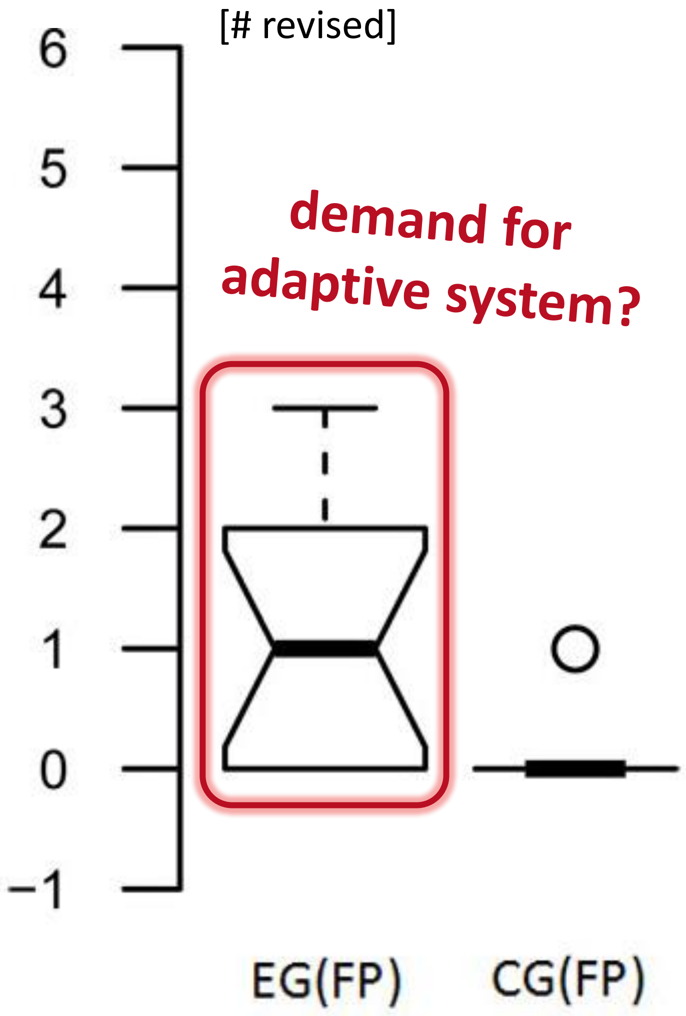
$\bar{x}_{CG(FP)} = 0.18$ (SE = 0.12)

Test statistic:

$t_{H2} = 2.55$

$|t_{H2}| > 2.06$ ($P = 0.017$)

- reject null hypothesis at 5% level
- significant difference b/w groups ✓



demand for adaptive system?

H3: Incomplete Feedback causes users to miss similar issues

Expectation: $\mu_{EG(FN)} \neq \mu_{CG(FN)}$

Arithmetic mean:

$\overline{x}_{EG(FN)} = 0.73$ (SE = 0.28)

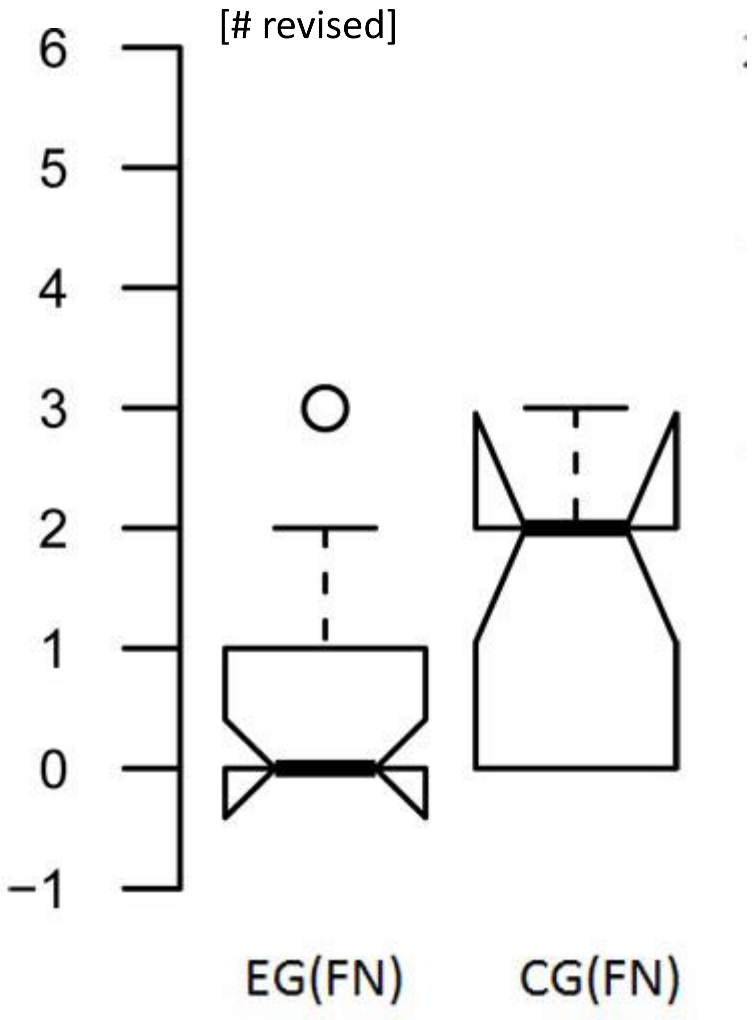
$\overline{x}_{CG(FN)} = 1.36$ (SE = 0.36)

Test statistic:

$t_{H3} = -1.39$

$|t_{H3}| < 2.06$ ($P = 0.17$)

- cannot reject null hypothesis
- cannot find significant difference ✗



H4: Task completion time similar

Expectation: $\mu_{EG(\tau)} = \mu_{CG(\tau)}$

Arithmetic mean:

$\bar{x}_{EG(\tau)}$ = 13 min, 3 sec (SE = 104 sec)

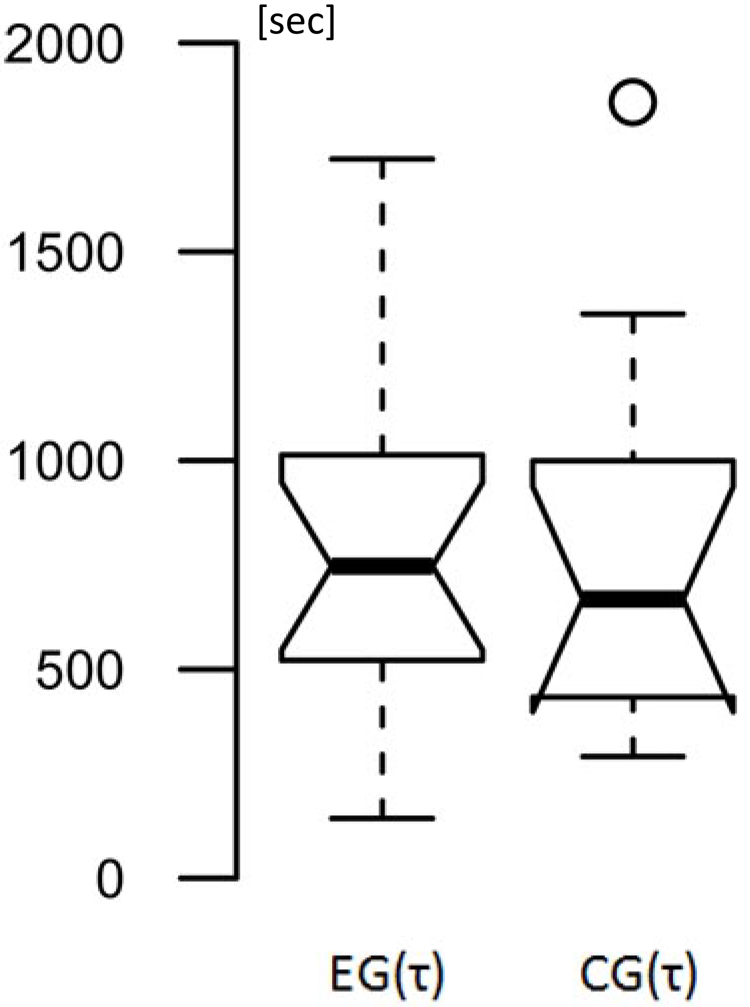
$\bar{x}_{CG(\tau)}$ = 13 min, 27 sec (SE = 144 sec)

Test statistic:

$t_{H4} = -0.14$

$|t_{H4}| < 2.06 \quad (P = 0.89)$

- cannot reject null hypothesis
- cannot find significant difference ✓



Conclusion

What we learned from this work:

- correct feedback helps (H1)
- incorrect feedback problematic (H2) – overtrust?
  - but: demand for **adaptive systems**!
- tendency to miss FNs, but not significant (H3)
  - confirms "**precision** more important than recall"
- using feedback didn't take longer (H4)

Software for evaluating writing assistance tools:
https://github.com/UKPLab/naacl-bea2016-writing-study

https://pixabay.com/en/hand-leave-pen-paper-thank-you-226358/ (CC0 Public Domain)

Kontakt / Contact

Christian M. Meyer
Technische Universität Darmstadt
Ubiquitous Knowledge Processing Lab

📫 Hochschulstr. 10, 64289 Darmstadt, Germany
☎ +49 (0)6151 16–25293
🖶 +49 (0)6151 16–25295
✉ meyer (at) ukp.informatik.tu-darmstadt.de

Rechtliche Hinweise

Die Folien sind für den persönlichen Gebrauch der Vortragsteilnehmer gedacht. Im Vortrag verwendete Photographien, Illustrationen, Wort- und Bildmarken sind Eigentum der jeweiligen Rechteinhaber oder Lizenzgeber. Um Missverständnisse zu vermeiden, wäre eine kurze Kontaktaufnahme vor Weitergabe oder -nutzung der Vortragsmaterialien empfehlenswert. Sofern Sie Ihre Rechte verletzt sehen, bitte ich ebenfalls um Kontaktaufnahme zur Klärung der Sachlage.

Legal Issues

The slides are intended for personal use by the audience of the talk. Photographies, illustrations, tradedmarks, or logos are property of the holder of rights. To avoid any misconceptions, I would strongly recommend to get in touch before reusing or redistributing the slides or any additional material of the talk. The same applies if you consider your rights infringed – please let me know to initiate further clarification.