

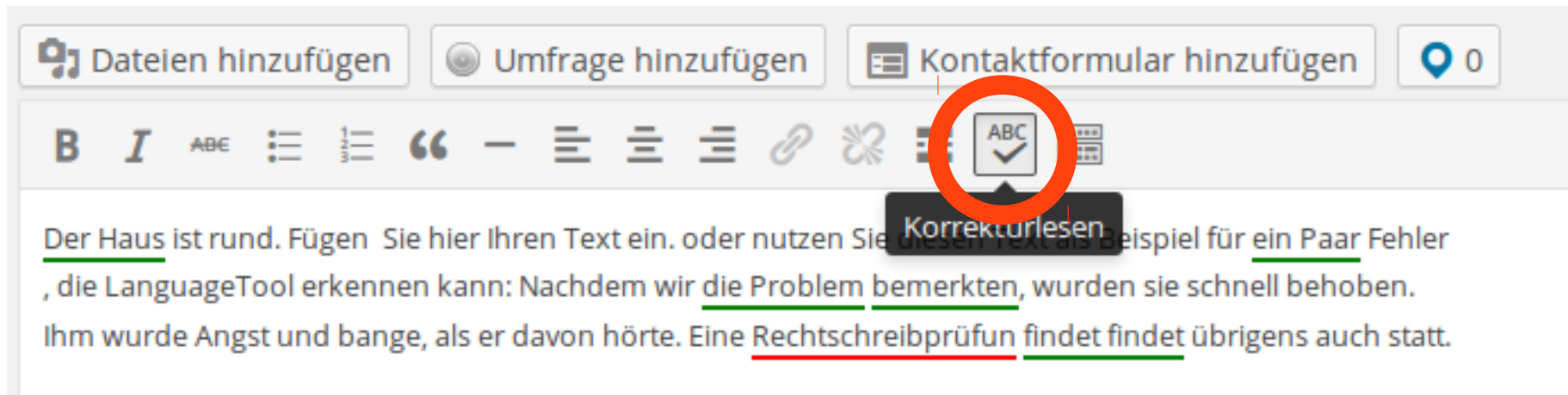
# Wie WordPress unsere Rechtschreibfehler findet



Daniel Naber  
WordCamp Hamburg 2014

# Ziel

- Diesen Button verstehen!




The screenshot shows a text editor interface. At the top, there are three buttons: 'Dateien hinzufügen', 'Umfrage hinzufügen', and 'Kontaktformular hinzufügen'. Below these is a rich text editor toolbar with various icons for bold, italic, text color, list, quote, link, and other formatting options. A red circle highlights the 'Korrekturlesen' button, which features an 'ABC' icon and a checkmark. A tooltip with the text 'Korrekturlesen' is visible below the button. The main text area contains the following text: 'Der Haus ist rund. Fügen Sie hier Ihren Text ein. oder nutzen Sie diesen Text als Beispiel für ein Paar Fehler, die LanguageTool erkennen kann: Nachdem wir die Problem bemerkten, wurden sie schnell behoben. Ihm wurde Angst und bange, als er davon hörte. Eine Rechtschreibprüfun findet findet übrigens auch statt.' The text contains several underlined words: 'Haus' (green), 'Fehler' (green), 'Problem' (green), 'bemerken' (green), 'Rechtschreibprüfun' (red), and 'findet' (green).



# Finde den Fehler

»Der Ticketverkauf für's  
WordCamp Hamburg ist ab  
sofort eröffnet.«



»Der Ticketverkauf ~~für's~~ fürs  
WordCamp Hamburg ist ab  
sofort eröffnet.«



# Finde den Fehler

»Ihr seid super! Auf  
deutsch gesagt AWESOME :D«




»Ihr seid super! Auf  
~~deutsch~~ Deutsch gesagt  
AWESOME :D«



# Finde den Fehler

»In diesem Jahr suchen wir nicht nur „normale“ Vorträge sondern auch kürzere, längere und interaktive Formate.«



»In diesem Jahr suchen wir nicht nur „normale“ Vorträge, sondern auch kürzere, längere und interaktive Formate.«

Man beachte: das sind mehr als nur Rechtschreibfehler - eine normale Rechtschreibkorrektur findet solche Fehler nicht





# Hilfe, Fehler! Was nun?

- Weitermachen wie bisher
- Wir werden alle Sprachexperten!
- Gibt es da nicht Hilfe in der Cloud?



📎 Dateien hinzufügen   🗳️ Umfrage hinzufügen   📄 Kontaktformular hinzufügen   📍 0

**B** *I* ABC ☰ ☷ ☹️ ☶ ☷ ☷ 🔗 🔁 ☰ ☑️ ☰

Der Haus ist rund. Fügen Sie hier Ihren Text ein, oder nutzen Sie diesen Text als Beispiel für ein Paar Fehler, die LanguageTool erkennen kann: Nachdem wir die Problem bemerkten, wurden sie schnell behoben. Ihm wurde Angst und bange, als er davon hörte. Eine Rechtschreibprüfun findet findet übrigens auch statt.

Korrekturlesen

# Wie funktioniert das?

- After the Deadline
  - Läuft als Service bei wordpress.com (hoher Speicherverbrauch)
  - Open Source
- Jetpack
  - Zugriff auf wordpress.com-Service per HTTPS
- Nutzt für nicht-englische Texte intern LanguageTool
  - auch Open Source



# Eine (sehr) kurze Geschichte der Textprüfung in WordPress

- 2009: Automattic kauft After the Deadline
- ca. 2011: After the Deadline wird nicht weiterentwickelt



# Aber!

- LanguageTool wird aktiv weiterentwickelt



# Wie wird ein deutscher Text geprüft?

- After the Deadline prüft Rechtschreibung der Wörter
- After the Deadline ruft LanguageTool auf
- LanguageTool:
  - zerlegt Text in Sätze
  - weist Wörtern ihre Wortart zu ('Häusern' = Nomen, Neutrum, Plural, Dativ)
  - sucht nach > 1800 Fehlermustern auf dem so analysierten Text
- After the Deadline filtert LanguageTool-Treffer mit Statistik der Worthäufigkeit

# Fehlermuster 1 (LanguageTool)

Case-sensitive word matching

⇅ Token #1

Word  Part-of-speech  Word + Part-of-speech  Any word

Word:   RegExp [?]  Base form  Negate

[Add exception](#) · [Edit Advanced Attributes](#) (0)

⇅ Token #2

Word  Part-of-speech  Word + Part-of-speech  Any word

Word:   RegExp [?]  Base form  Negate

[Add exception](#) · [Edit Advanced Attributes](#) (0)

# Fehlermuster 2 (LanguageTool)

Case-sensitive word matching

⇅ Token #1

Word  Part-of-speech  Word + Part-of-speech  Any word

Word: auf

RegExp [?]  Base form  Negate

[Add exception](#) · [Edit Advanced Attributes](#) (0)

⇅ Token #2

Word  Part-of-speech  Word + Part-of-speech  Any word

Word: englisch|französisch|deutsch

RegExp [?]  Base form  Negate

[Add exception](#) · [Edit Advanced Attributes](#) (0)



# Fehlermuster 3 (LanguageTool)

↕ Token #2 ✕

Word  Part-of-speech  Word + Part-of-speech  Any word

Part-of-speech:   RegExp [?]  Negate

[Add exception](#) · [Edit Advanced Attributes](#) (0)

## Part-of-Speech Help ✕

1: 1. Person · 2: 2. Person · 3: 3. Person · **SUB: Substantiv/Nomen** · EIG: Eigennamen · VER: Verb · ADJ: Adjektiv · ART: Artikel · PRO: Pronomen · ADV: Adverb · PRP: Präposition · NEG: Negationspartikel · ABK: Abkürzung · AKK: Akkusativ · AUX: Hilfsverb · BEG: begleitend · B/S: begleitend oder stellvertretend · CAU: kausal · DAT: Dativ · DEF: bestimmt · DEM: Demonstrativpronomen · EIZ: erweiterter Infinitiv mit zu · FEM: femininum · GEN: Genitiv · GRU: Grundform · IND: unbestimmt · INF: Infinitiv · IMP: Imperativ · INR: Interrogativpronomen · KJ1: Konjunktiv: 1 · KJ2: Konjunktiv: 2 · KOM: Komparativ · KON: Konjunktion · LOK: lokal · MAS: maskulinum · MOD: modal · NEB: nebenordnend · NEU: neutrum · NOA: ohne Artikel · NOG: ohne Genus · NOM: Nominativ · NON: nicht-schwach · PA1: Partizip 1 · PA2: Partizip 2 · PER: personal · **PLU: Plural** · POS: possessiv · PRÄ: Präsens · PRD: prädikativ · PRI: proportional · PRT: Präteritum, Imperfekt · REF: reflexiv · RIN: relativ oder interrogativ · SFT: schwach · SIN: Singular · SOL: alleinstehend · STV: stellvertretend · SUP: Superlativ · TMP: temporal · UNT: unterordnend · VGL: vergleichend · ZAL: Zahlwort · ZUS: Verbzusatz

# Filtern der Fehlermeldungen mit Statistik (After the Deadline)

- "auf [Dd]eutsch gesagt"?
  - auf deutsch: 30 Vorkommen
  - auf Deutsch: 35 Vorkommen
  - deutsch gesagt: 10 Vorkommen
  - Deutsch gesagt: 10 Vorkommen  
(das sind nur Beispielzahlen)

# Filtern der Fehlermeldungen mit Statistik (After the Deadline)

- "auf **d**eutsch gesagt":  $30+10=40$
- "auf **D**eutsch gesagt":  $35+10=45$
- Vorkommen in der Wikipedia (ca. 2010), Blogs
  - Problem, wenn zu wenige Vorkommen für beide Varianten



# Zusammenfassung

- WordPress nutzt After the Deadline als Service zur Textprüfung
- After the Deadline nutzt LanguageTool für nicht-englische Texte
- LanguageTool kennt > 1800 Fehlerregeln, nach denen es den Text durchsucht
- LanguageTool-Ergebnis wird von After the Deadline nochmal gefiltert

# Und nun?

- Ausprobieren!
  - wordpress.com
  - Jetpack
  - <https://languagetool.org>
- Fehler an mich :)
- Wikipedia-Änderungen prüfen:  
<http://community.languagetool.org/feedMatches/>

# Wikipedia verbessern

- <http://community.languagetool.org/feedMatches/>

## Prüfung von Wikipedias 'Letzte Änderungen' (785)

Hinweis: die Rechtschreibprüfung ist für diese Prüfung der 'Letzten Änderungen' nicht aktiv

Fehler, die seit 24 Stunden bestehen ▾ - alle Kategorien - ▾

- alle nicht-versteckten Regeln - ▾



Änderungsdatum	Treffer
<b>2014-06-03</b>	
2014-06-03 09:31	<b>Fügen Sie zwischen Sätzen ein Leerzeichen ein</b> Sonstiges &lt;ref&gt;Volksbelustigungen von Dr. <u>Florian</u> Dering ISBN 9783891900055&lt;/ref&gt; Seite jetzt prüfen · Änderungen · Fritz Hilbert
2014-06-03 09:04	... sh. z.B. <a href="https://fr.wikipedia.org/wiki/Louis_de_Silvestre#.C5.92uvres">https://fr.wikipedia.org/wiki/Louis_de_Silvestre#.C5.92uvres</a> --&gt; Seite jetzt prüfen · Änderungen · Louis de Silvestre
2014-06-03 08:30	...eröffnetlicht wurde.&lt;ref&gt; <a href="https://www.usp.gv.at/Portal.Node/usp/public?genetics.am=Content&amp;amp;p.contentid=10007.64383&amp;l...">https://www.usp.gv.at/Portal.Node/usp/public?genetics.am=Content&amp;amp;p.contentid=10007.64383&amp;l...</a> Seite jetzt prüfen · Änderungen · Weltraumregistrierungsübereinkommen
2014-06-03 08:21	Er wurde am 02. <u>Juni</u> 2014, knapp ein halbes Jahr vorm eigentlichen Termin, entla... Seite jetzt prüfen · Änderungen · George Jung
2014-06-03 08:19	<b>Vor der Konjunktion 'sondern' steht immer ein Komma.</b> Zeichensetzung ...rem das UN-Hochkommissariat für Flüchtlinge (UNHCR), würden <u>nicht objektiv sondern</u> mit "doppelten Standards" arbeiten. Es gehe diesen Institut... Seite jetzt prüfen · Änderungen · Krise in der Ukraine 2014
2014-06-03 08:04	<b>Fügen Sie zwischen Sätzen ein Leerzeichen ein</b> Sonstiges ...weiteren Verbleib des Vlies beschreiben, sind nicht bekannt. <u>Aber</u> alles ist nicht wahr meint pa. Seite jetzt prüfen · Änderungen · Goldenes Vlies

Thank you and have a **niece** conference!

Did you mean 'nice' (=pleasant)?

nice

Ignore this error

### Zum Nachlesen:

Raphael Mudge: The Design of a Proofreading Software Service

<http://www.aclweb.org/anthology/W/W10-0404.pdf>

Daniel Naber: A Rule-Based Style and Grammar Checker

[http://danielnaber.de/language-tool/download/style\\_and\\_grammar\\_checker.pdf](http://danielnaber.de/language-tool/download/style_and_grammar_checker.pdf)



This presentation is licensed under CC-BY 4.0 <http://creativecommons.org/licenses/by/4.0/>