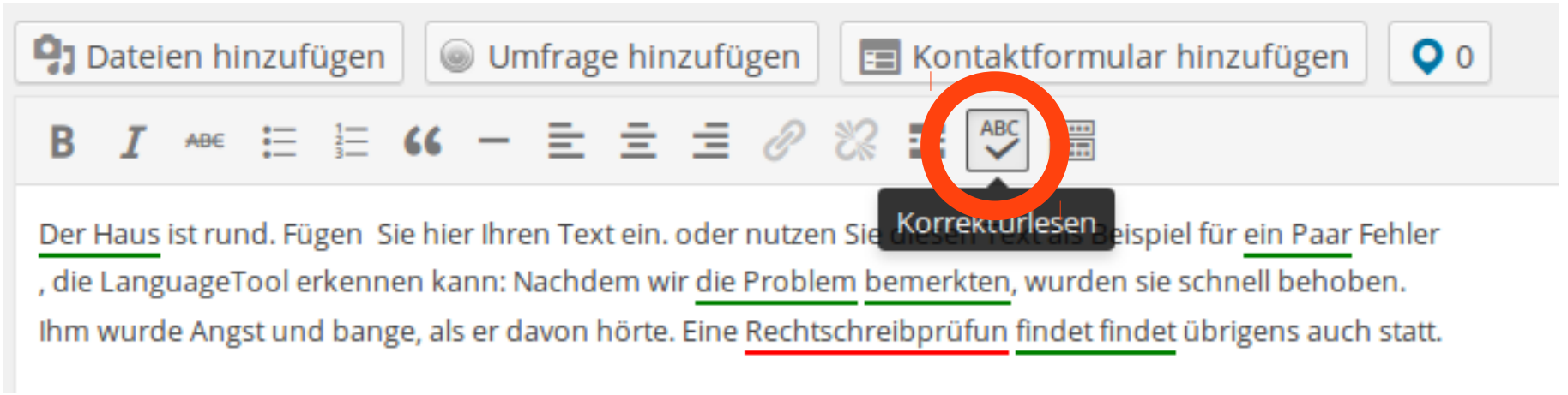
Wie WordPress unsere Rechtschreibfehler findet

Daniel Naber
WordCamp Hamburg 2014

Ziel

- Diesen Button verstehen!



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

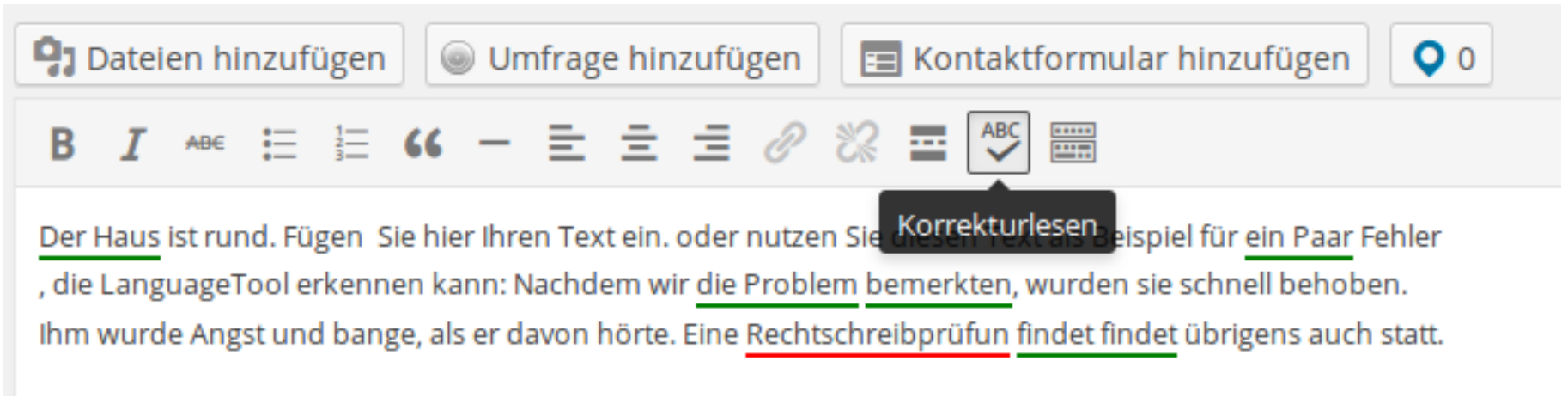
»In diesem Jahr suchen wir nicht nur „normale“ Vorträge sondern auch kürzere, längere und interaktive Formate.«

»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
ABC
Korrekturlesen
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.

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

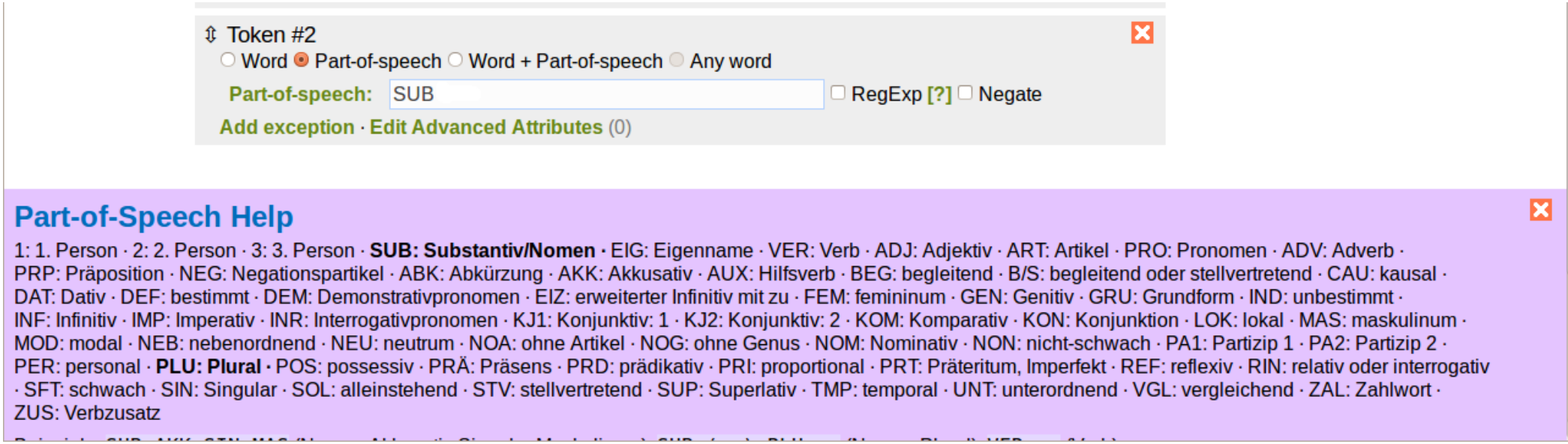
Fehlermuster 1 (LanguageTool)

http://community.languagetool.org/ruleEditor2/

Fehlermuster 2 (LanguageTool)

http://community.languagetool.org/ruleEditor2/

Fehlermuster 3 (LanguageTool)



http://community.languagetool.org/ruleEditor2/

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