Automatische Prüfung von Wikipedia-Artikeln

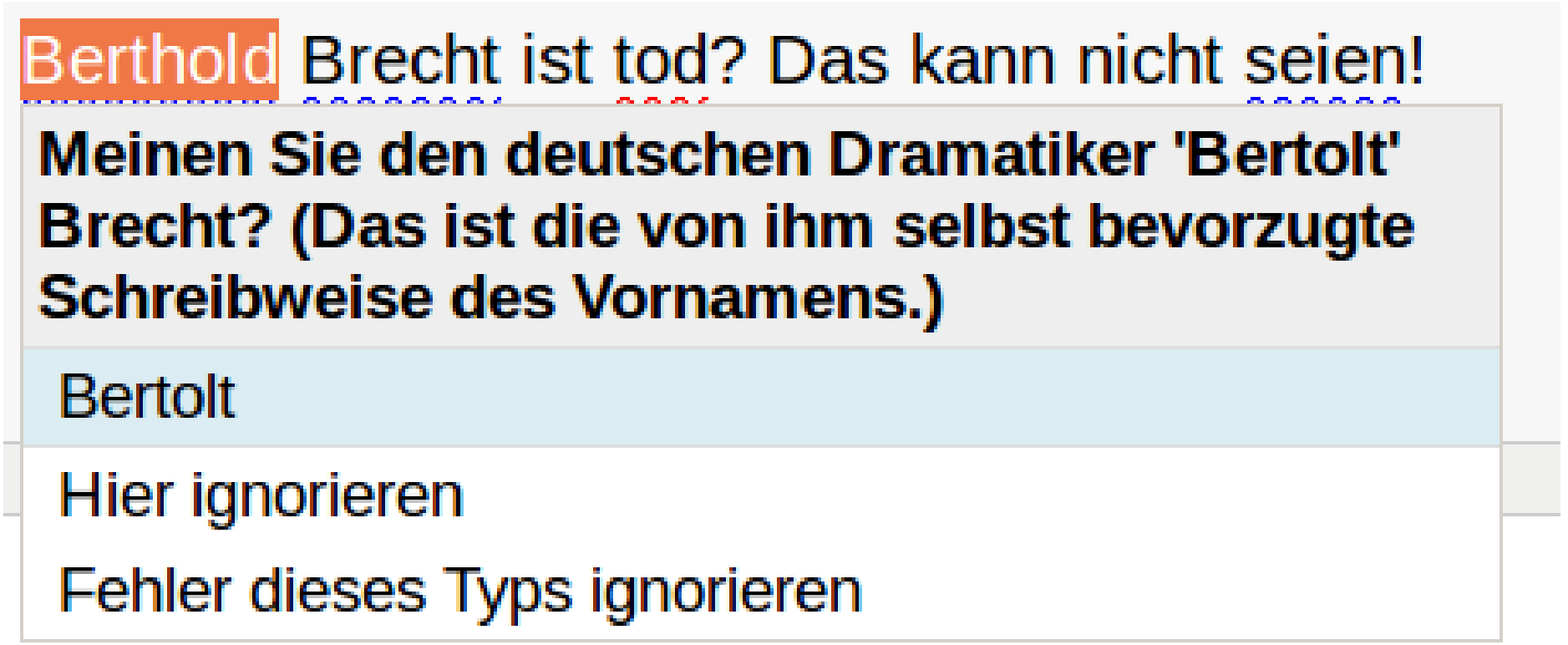
Wikimedia Deutschland - Offener Sonntag
2013-05-26

Daniel Naber

Berthold Brecht ist tod?
Das kann nicht seien!

(Finde die Fehler)

Berthold Brecht ist tod? Das kann nicht seien!


Berthold Brecht ist tod? Das kann nicht seien!
Meinen Sie den deutschen Dramatiker 'Bertolt' Brecht? (Das ist die von ihm selbst bevorzugte Schreibweise des Vornamens.)
Bertolt
Hier ignorieren
Fehler dieses Typs ignorieren

Über

- http://www.danielnaber.de
- http://wikifeedme.openthesaurus.de

Lieblingsfehler

- Seid gestern
- größer wie
- gemäß des
- Mühlheim an der Ruhr
- Siegmund Freud
- spiegelt ... wieder
- "Ein Zitat"
- Mittelohrenzündung
- Email
- Kommuniqué

Lieblingsfehler 2

- Seid gestern > Seit gestern
- größer wie > größer als
- gemäß des > gemäß dem
- Mühlheim an der Ruhr > Mülheim an der Ruhr
- Siegmund Freud > Sigmund Freud
- spiegelt ... wieder > spiegelt ... wider
- "Ein Zitat" > „Ein Zitat“
- Mittelohrenzündung (Mittel + ohren + zündung) > Mittelohrentzündung
- Email > E-Mail
- Kommuniqué/Kommunikee

Wikipedia-spezifisches

- Seit Kurzem > relative Zeitangaben
- Wikipedia:Vermeide hohle Phrasen
  - "entgegen landläufiger Meinung"
  - "Die Wissenschaft sagt"

LanguageTool

- Open Source Stil- und Grammatikprüfung
- Lizenz: LGPL
- erste Version 2003
- unterstützt jetzt 29 Sprachen
- Homepage: http://www.languagetool.org

LanguageTool 2

- LanguageTool fängt da an, wo die klassische Rechtschreibprüfung aufhört
- klassische Rechtschreibprüfung = prüft einzelne Wörter, ohne Kontext
- Rechtschreibprüfung: schon im Browser vorhanden, deshalb hier kein Thema

LanguageTool-Prinzip

- Text wird analysiert: Erkennung Satzgrenzen, Wortarten
- Text wird auf Fehlermuster durchsucht
  - LanguageTool kennt keine Grammatik, es kennt Fehler
- Komplizierte Regeln in Java programmiert
- Warum war "Brecht" falsch im ersten Beispiel?

Berthold Brecht ist tod? Das kann nicht seien!

Fehlermuster-Beispiel (vereinfacht)

- <rule>

  <pattern>
  <token>Berthold</token>

  <token>Brecht</token>
  </pattern>

  <message>Meinen Sie den deutschen Dramatiker
  <suggestion>Bertolt Brecht</suggestion>?</message>

  </rule>

Fehlermuster

- Reguläre Ausdrücke:
  <token regexp="yes">Berth?old</token>
- Wortarten ansprechen, z.B. alle Verben:
  <token postag_regexp="yes" postag="VER:.*">
- Negation:
  <token negate="yes">Haus</token>
- URL:
  <url>http://www.korrekturen.de/wortliste/das_wenige.shtml</url>
- ..und mehr: http://www.languagetool.org/development/

Fehlermuster 2

- Deutsch: über 1700 dieser Regeln
  - Französisch, Katalanisch, Polnisch:
    > 1000 Regeln
  - Englisch: 900 Regeln
  - Griechisch, Japanisch, Khmer: < 25 Regeln

Regeln selber schreiben

- Rule creator:
  http://www.languagetool.org/ruleeditor/
  - prüft gegen 10.000 Wikipedia-Seiten
  - Advanced mode
    Beispiel/Testsätze im XML

LanguageTool und Wikipedia

- Fehlersammlung:
  http://community.languagetool.org/
  (ca. 5000 Artikel aus dem XML-Dump)
- Einzelne Seite prüfen:
  http://www.languagetool.org/wikicheck/
  auch mit Bookmarklet
- HTTP-Schnittstelle für Entwickler

TODOs

- Firefox-Plugin - kennt Mediawiki-Syntax nicht und arbeitet auch noch nicht mit dem VisualEditor zusammen
- mehr Wikipedia-spezifische Regeln
- ...

Vielen Dank!