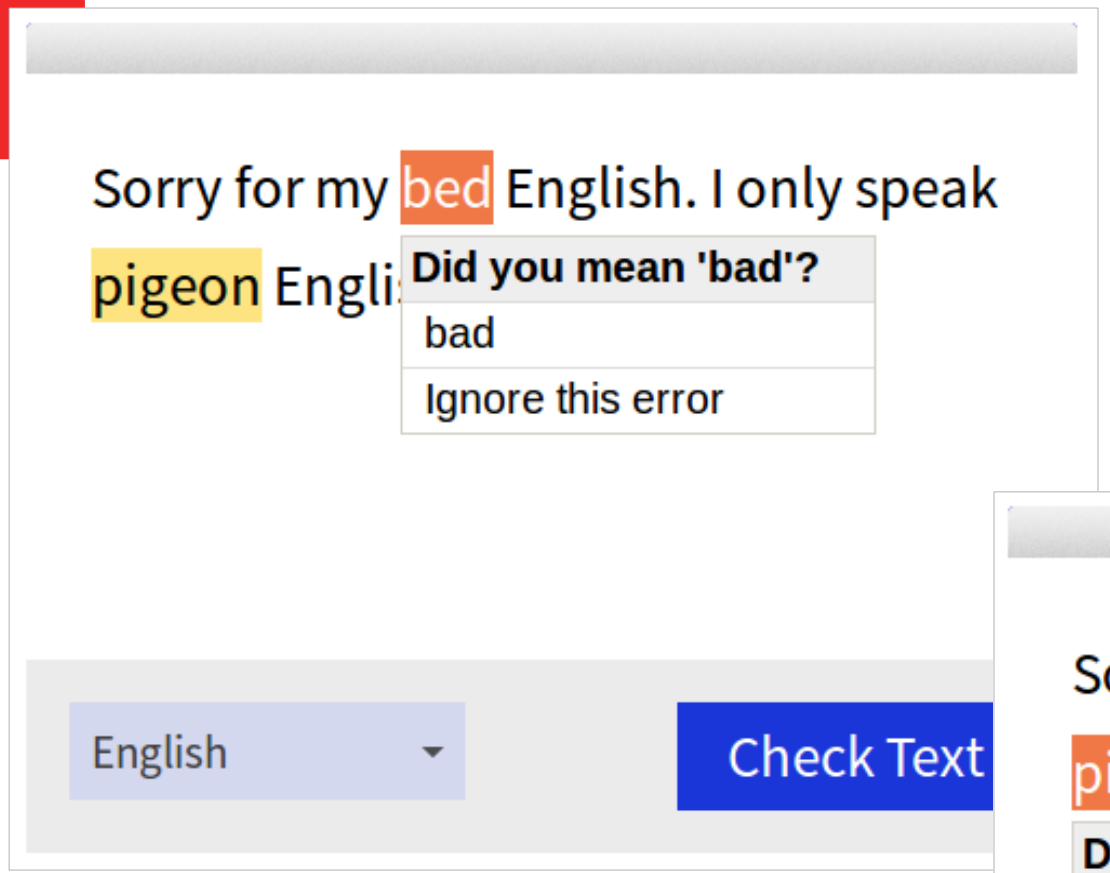
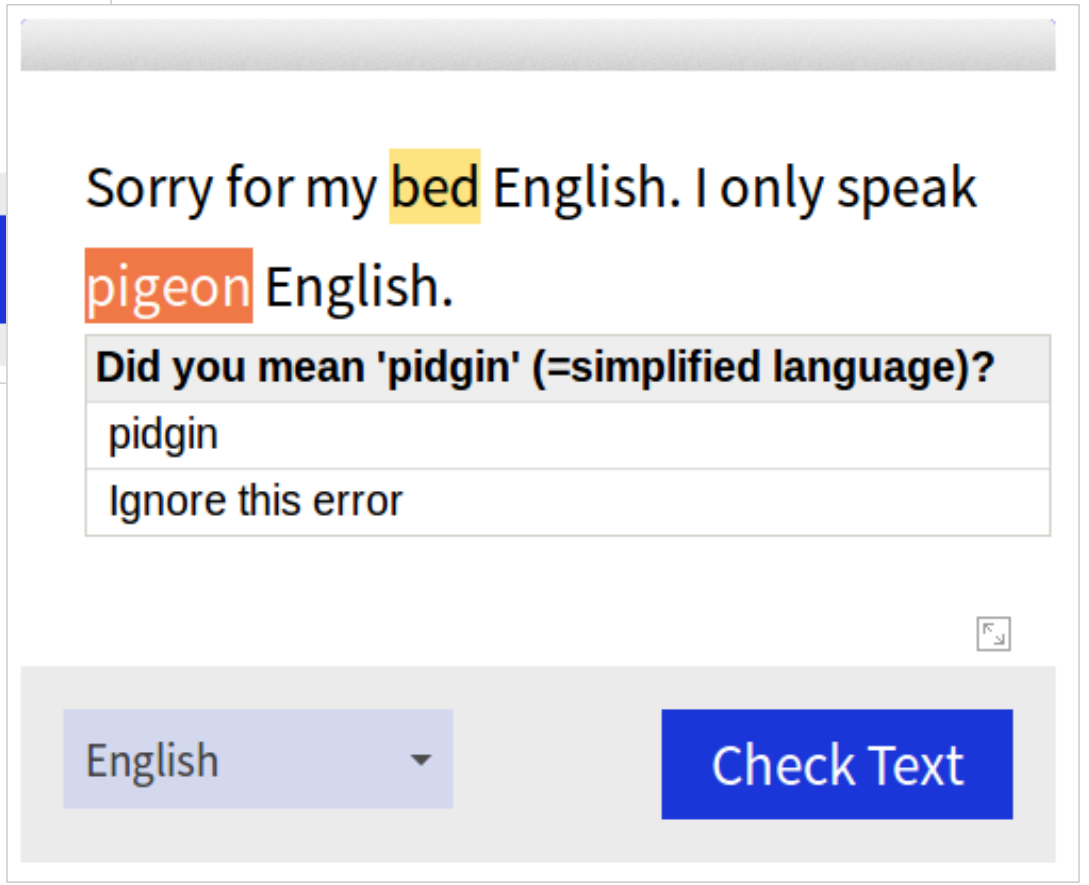
How we found a million style and grammar errors in the English Wikipedia...
and how to fix them

Daniel Naber
FOSDEM 2014

- Sorry for my bed English
- I only speak pigeon English

Sorry for my ~~bed~~ bad English


Sorry for my bed English. I only speak
pigeon Engli
Did you mean 'bad'?
bad
Ignore this error
English
Check Text


Sorry for my bed English. I only speak
pigeon English.
Did you mean 'pidgin' (=simplified language)?
pidgin
Ignore this error
English
Check Text

Roadmap

- How did we find one million errors in Wikipedia?
- How does LanguageTool work?
- Why not use a different approach?
- How to fix the million errors?
- Future work

Survey

- How many people here have heard of LanguageTool?
- How many people have used it?

How to find one million errors in Wikipedia

- `java -jar languagetool-wikipedia.jar check-data -f enwiki-20140102-pages-articles.xml -l en`
  - enwiki-20140102-pages-articles.xml = Wikipedia XML dump
  - en = language code for English

How to find one million errors in Wikipedia: Output

- Title: Alabama

  1.) Line 1, column 47

  Message: The verb 'will' requires base form of the verb: 'designate'.

  A proposed northern bypass of Birmingham will designated as I-422.
                          ^^^^^^^^^^

How to find one million errors in Wikipedia (cont.)

- Run on 20,000 articles
  - Takes about 10ms per sentence (English)
- Got 37,000 potential errors
  - Error: grammar error, style suggestions
- Projection to the whole Wikipedia (4.4m articles): 8 million potential errors
- Checked about 200 randomly selected potential errors manually
- Result: 1 million errors
  - Not counting errors from a simple spell checker

LT

Why so many false alarms?

- Difficult text extraction from Wikipedia
  - Mediawiki syntax, e.g. templates not expanded:
    "an elevation of about {{convert|115|m|ft}}"
- Many non-English names, places, movie titles, …
- Articles about math:
  "The value of n for a given a is called …"
- Articles have been checked already
- Our English rules need to be improved

Examples: Bad matches

- … and 68000 assembler …
  - Suggestion: assemblers
- Score voting and Majority Judgment allow these voters …
  - Suggestion: allows
- If a is algebraic over K
  - Suggestion: an

Examples: Useful matches

- In a vote of 27 journalists from 22 gaming magazine, …
  - Suggestion: magazines
- An energy called qi flows through through the body …
  - Suggestion: through
- … sending back their work to the teachers computer.
  - Suggestion: teacher's, teachers'

Examples: Style

- ... but there are many different variations.
  - Suggestion: many

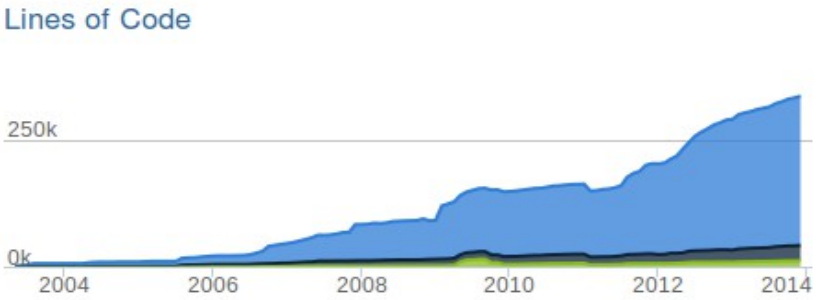
Examples: Errors not detected

(not from Wikipedia)

- Sematic problems: “Barack Obama is the president of France”
- “I made a concerted effort.”
- Tenses: “Tomorrow, I go shopping.”

LanguageTool Overview

- Idea: the next step after spell checking
- Started in 2003
- LGPL
- About 10 regular committers
- New release every 3 month
- Implemented in Java + XML



How to use LanguageTool?

- As a command-line application and desktop application
- As an extension:
  - LibreOffice/OpenOffice
  - Vim, Emacs
  - Firefox, Thunderbird
- As a Java API
- Via HTTP, returns simple XML
  - comes with an embedded HTTP server

How does LanguageTool work?

1. Takes plain text as input
2. Splits text into sentences
3. Splits sentences into words
4. Finds part-of-speech tags for each word and its base form (walks → walk)
5. Matches the analyzed sentences against error patterns and runs Java rules

Error detection patterns

- Patterns make it easy to contribute to LanguageTool: no programming needed & no dependencies between patterns
- Slightly simplified example:

```
<rule>
    <pattern>
        <token>bed</token>
        <token regexp="yes">English|attitude</token>
    </pattern>
    <message>
     Did you mean <suggestion>bad \2</suggestion>?
    </message>
</rule>
```

Error detection patterns (cont.)

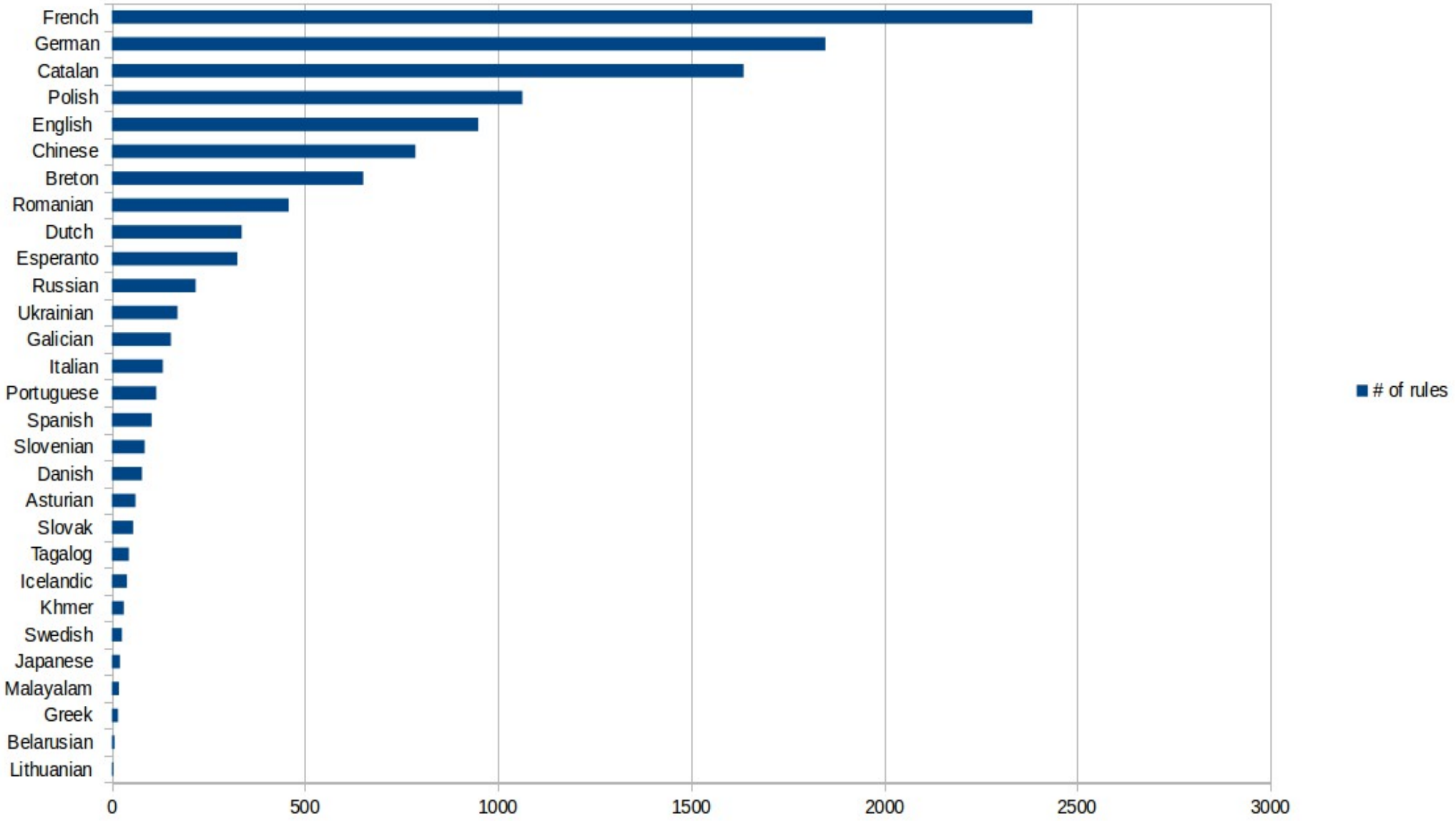
- Pattern features
  - Logical OR, AND
  - Negation
  - Skipping
  - Inflection
  - Match part-of-speech
  - See http://wiki.languagetool.org/development-overview

Error detection patterns (cont.)

```
<rule>
<pattern>
   <token postag="SENT_START"/>
   <token regexp="yes">Always|Hardly|Never</token>
   <token><exception postag="VB.*|MD|JJ"
                              postag_regexp="yes"/></token>
</pattern>

<message>The adverb '\2' is usually not used at the beginning of a
              sentence.</message>

<example type="incorrect">Always I am happy.</example>
<example type="correct">I am always happy.</example>

</rule>
```

Error detection patterns (cont.)

- Support for 29 languages (to a very different degree)



- Why not use a more powerful approach?

What is grammar?

- Grammar is a set of rules that describe how valid words, sentences, and texts look like
- Syntax is a formal description of how a valid sentence looks like
- What is a parser?
  - Takes an input sequence and creates a structure, e.g. a tree
  - This is similar for natural languages and programming languages, so...

LT

So why not develop a parser for English?

- It's difficult, as English wasn't made for being parsed
  - "spec" about 1700 pages ("A Comprehensive Grammar of the English Language")
  - "spec" about 700 pages (Esperanto, "Plena Manlibro de Esperanta Gramatiko")
- It would be mostly specific to English

So why not develop a parser for English? (cont.)

- Parser != good error messages
- You'll need rules anyway - “Sorry for my bed English” parses fine
- There are parsers, though (e.g. Link Grammar)

Why not use machine learning?

- We *do* use OpenNLP for chunking
- You'd probably need an error corpus
- But feel free to do that, just implement your own rule in Java

LT

When error patterns are not enough

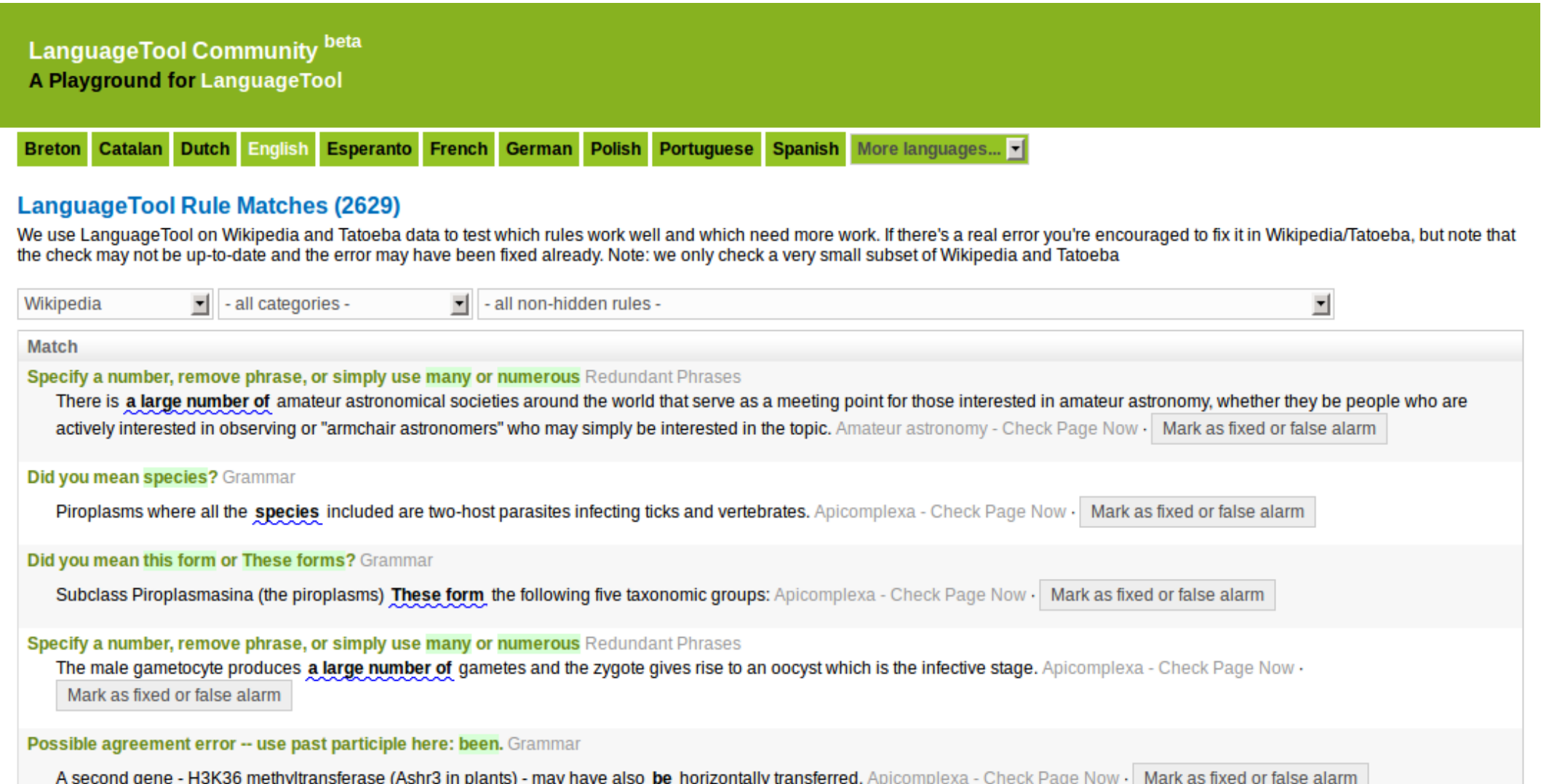
- implement Rule.match()

```
@Override
public RuleMatch[] match(AnalyzedSentence as) {
  AnalyzedTokenReadings[] tokens = as.getTokens();
  // find errors here
}
```

How to fix the million Wikipedia errors?

How to fix the million Wikipedia errors?

- You could look at the mass check and fix errors, but...
  http://community.languagetool.org/corpusMatch



How to fix the million Wikipedia errors? (cont.)

- Fix errors from the 'Recent Changes' feed check http://community.languagetool.org/feedMatches
- Fetches the Atom Feed of changes about twice a minute
- Checks only the parts that have been modified
- Detects if an error gets fixed

How to fix the million Wikipedia errors? (cont.)

Check of Wikipedia's Recent Changes (15820)

Please note that spell checking is not activated yet for this 'Recent Changes' check

Errors not fixed for 24 hours | - all categories -

- all non-hidden rules -

Edit Date	Match
2014-01-13	
2014-01-13 12:00	**Consider using a past participle here: criticised.** Grammar ...Me the Horizon album)|Count Your Blessings]], the album was **criticise** for its musical aesthetics. **Check Page Now** · Diff · Suicide Season · Mark as fixed or false alarm
2014-01-13 11:57	**Possible typo: you repeated a word** Miscellaneous '''Lasagne''' ({{IPAc-en|l|ə|'|z|æ|n|j|ə}} or **{{IPAc-en|l|ə|'|z|ɑ:|n|j|ə}} or en|l|ə|'|s|ɑ:|n|j|ə}}, {{IPA-it|la'zaɲɲe}}**, singular '''lasagna''') are a wid shape, and... **Check Page Now** · Diff · Lasagne · Mark as fixed or false alarm
2014-01-13 11:57	**Possible typo: apostrophe is missing. Did you mean Communications' or communication's?** Possible Typos ...ri who was a senior TV anchor and journalist is the current **Communications** A Prime Minister of India **Check Page Now** · Diff · NDTV India · Mark as fixed or false alarm

How to fix the million Wikipedia errors? (cont.)

Consider using a past participle here: criticised.

…izon album)|Count Your Blessings]], the album was **criticise** for its song writing and musical aesthetics…

select correction below **reset** **copy**

- **criticised**

Use past participle here: gathered.

…nd in drummer Matt Nicholls' opinion the band had **gather** strong hatred from 'proper [[Heavy metal sub…

select correction below **reset** **copy**

- **gathered**

Use an instead of 'a' if the following word starts with a vowel sound, e.g. 'an article', 'an hour'

…t in Arboga the band caused a controversy and lit **a** unlit, prepared bonfire in the middle of th…

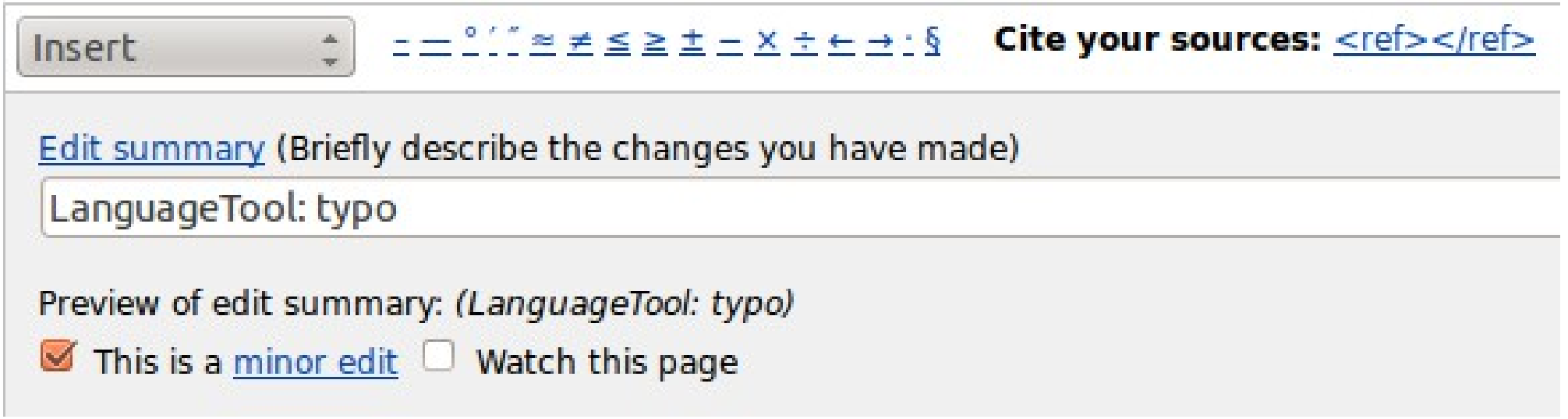
select correction below **reset** **copy**

- **an**

How to fix the million Wikipedia errors? (cont.)

− ''Suicide Season'' spawned four singles ("The Comedown", "Chelsea Smile", "Diamonds Aren't Forever" and "The Sadness Will Never End"). The album debuted on the charts of five countries. Critically the album received a mixed response. Though praised from the musical shift from the style of 2006's [[Count Your Blessings (Bring Me the Horizon album)|Count Your Blessings]], the album was **criticise** for its song writing and musical aesthetics.	+ ''Suicide Season'' spawned four singles ("The Comedown", "Chelsea Smile", "Diamonds Aren't Forever" and "The Sadness Will Never End"). The album debuted on the charts of five countries. Critically the album received a mixed response. Though praised from the musical shift from the style of 2006's [[Count Your Blessings (Bring Me the Horizon album)|Count Your Blessings]], the album was **criticised** for its song writing and musical aesthetics.
==Background and recording==	==Background and recording==

How to fix the million Wikipedia errors? (cont.)



Future Work

- Wish: make style and grammar checking ubiquitous (like spell checking already is)
- Current State
  - (+) stable Java API (on Maven Central), HTTP API
  - (+) support for many languages
  - (+) license (LGPL)
  - (+/-) Java
    - Solution? Compile to Javascript (LLVM)

Help Needed

- Compile Java to Javascript (LLVM)
  - http://stackoverflow.com/questions/19902556
- Add support for another language
- Need maintainers for: English, Belarusian, Chinese, Galician, Icelandic, Japanese, Lithuanian, Malayalam, Brazilian Portuguese, Romanian, Swedish, Danish

Summary

- No need to stick to spell checking today – more powerful checks are available
- Style and grammar checking is useful for finding errors in Wikipedia
- Your contributions are welcome

Thank you and have a niece conference!

Did you mean 'nice' (=pleasant)?
nice
Ignore this error

Homepage: https://languagetool.org
Source code: https://github.com/languagetool-org/languagetool