# *Tool-supported Linguistic Quality in Web-related Multilanguage Scenarios*

**Inna Nickel, SAP AG; Daniel Naber, LanguageTool; Christian Lieske, SAP AG**

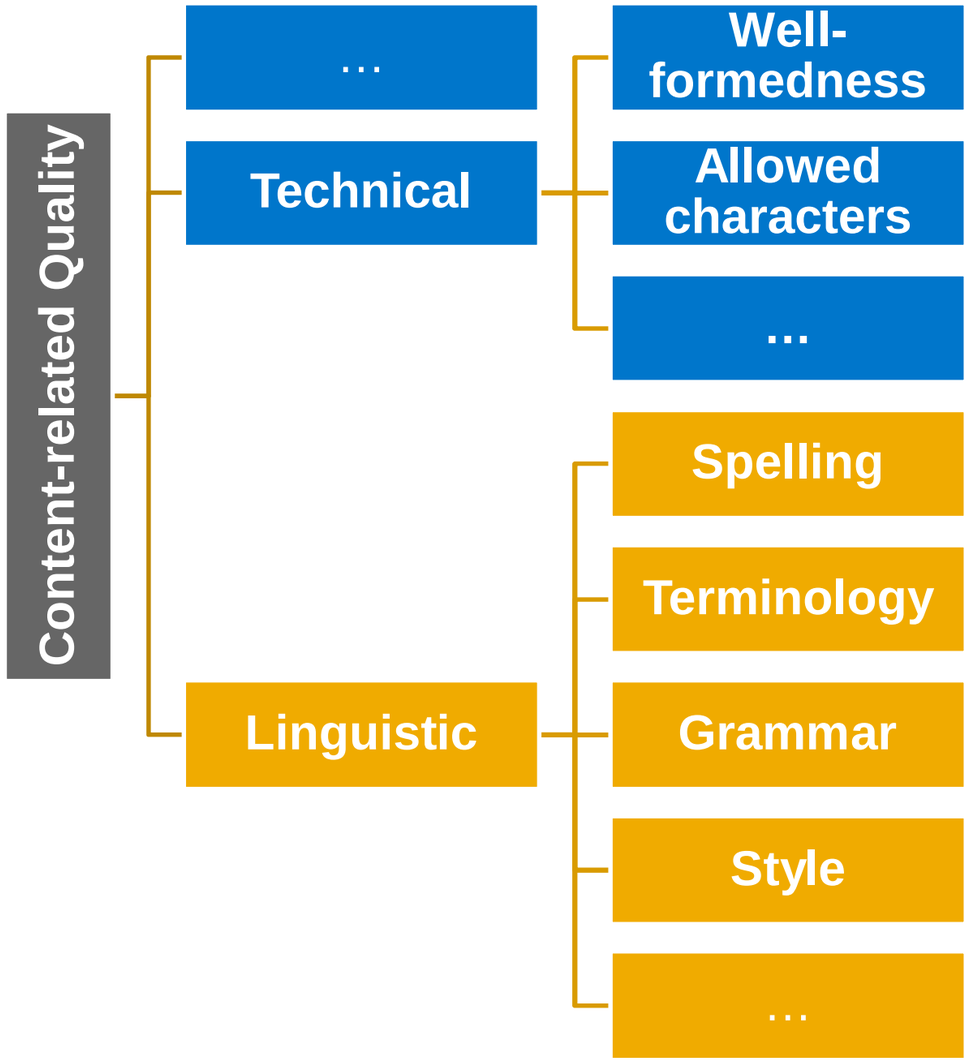
W3C Workshop
Making the Multilingual Web Work
12-13 March 2013, Rome


This presentation is
— our employers ha
for the work, tools, a
contained here

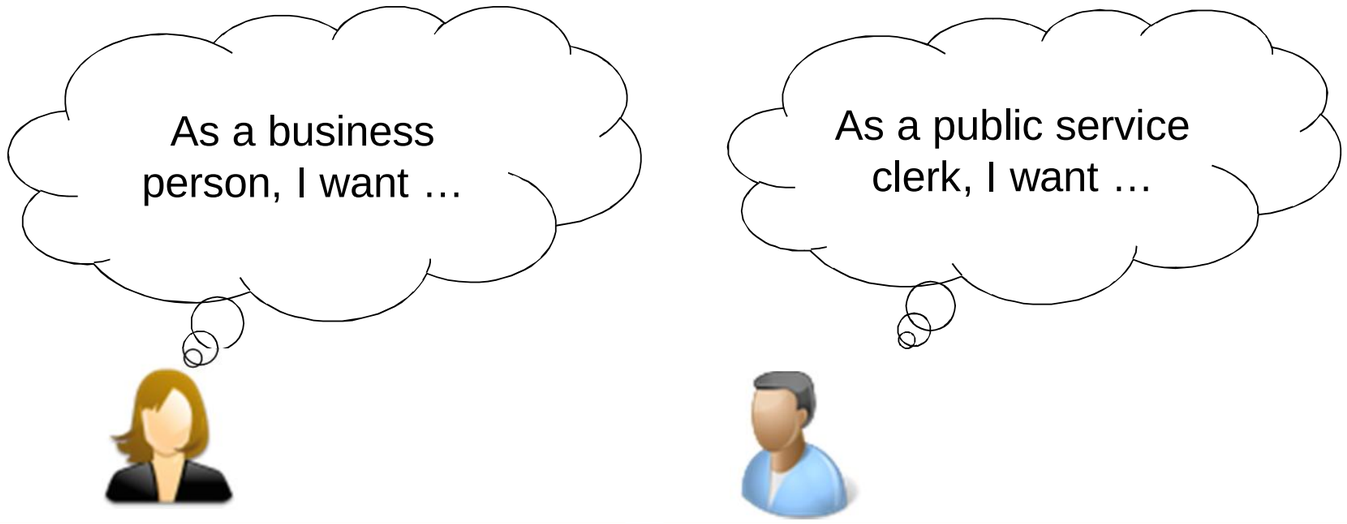
# Overview

- Scenario-dependant linguistic quality (SLQ)
- SLQ and the multilingual Web
- Natural Language Processing and Web-related SLQ
- An Open Source tool for SLQ
- Experiences from real-world deployments

# Scenario-dependant linguistic quality (1/3)



# Scenario-dependant linguistic quality (2/3)




- **My company's brand names and terminology**
- **Marketing speak with a people-centric/conversational style/voice**

1. "Make sure any offerings are properly trademarked"
2. Never include trademarks on Web sites (rather link) to our copyright/trademark site

- **Sober factual information**
- **Grammar that can be understood by 8th grade pupils**

1. Sentences with a single dependent clause (BITV 2.0 - German legislation)
2. Avoid genitives (Easy-to-Read guidelines)

# Scenario-dependant linguistic quality (3/3)

## Company/Business/Enterprise

### General guidelines for source language

**Latin Abbreviations**

1. Do **not** use Latin abbreviations. Instead, use the full English equivalent.

EXAMPLE

| ✔ (*Correct*) Full English Term | ✖ (*Incorrect*) Latin Abbreviation |
|---|---|
| for example | e.g. |
| that is | i.e. |
| and so on | etc. |
| note | n.b. |
| and others | et al. |
| opposed to, versus | vs. |
| compare | cf. |

End of: en-US

### Addendum for marketing

- Do not insert between common prefixes such as pre-, sub-, non-, mid-, or inter-
- Use with all “e-words,” such as “e-shopping” or “e-reader”
- Do not put spaces before or after a hyphen

### Guidelines for (translation into) Russian

**EXAMPLES**

| | |
|---|---|
| DE | Um das Programm aufzurufen, wählen Sie *Starten*. |
| EN | To call t the system, choose *Start*. |
| ✔ | Для вызова программы выберите *Запуск*. |
| ✖ | Программа вызывается при помощи *Запуска*. |

### Validated terminology

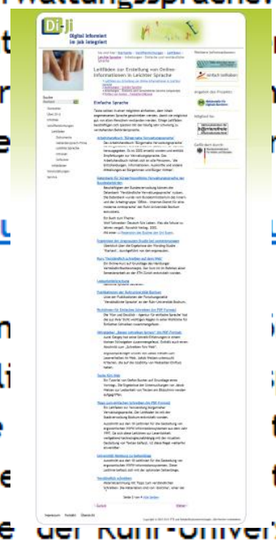
e-mail
электронная почта

e-mail
Invalid Synonym Of:
employee e-mail
email
почта

## Public Service

Arbeitshandbuch "Bürgernahe Verwaltungssprache"

Das Arbeitshandbuch "Bürgernahe Verwaltungssprache" wird vom Bundesverwaltungsamt - Bundesstelle für Büroorganisation und Bürotechnik (BBB) im PDF Format herausgegeben. Es ist 2002 erstellt worden und enthält Empfehlungen zur Verwaltungssprache. Das Arbeitshandbuch richt rsonen, "die Entscheidungen, Infor fte und andere Mitteilungen an Bürge r richten".



Datenbank für bürgerfreu ungssprache der Bundesbehörden

Beschäftigten der Bun nnen die Datenbank "Verständli prache" nutzen. Die Datenbank wurde terium des Innern und der Arbeitsgruppe t-Dienst für eine moderne Amtssprache der Ruhr-Universität Bochum entwickelt.

# Scenario-dependant linguistic quality and the multilingual Web


# Natural Language Processing and Web-related SLQ (1/2)

1. **Natural Language Processing (NLP) is for example the base of voice control, machine translation, and … linguistic quality control such as style checking**
2. **NLP systems usually require adaptation for a specific usage setting – for example may need to be “taught” about company-specific terminology**
3. **Adaptation either means that linguistic knowledge is formalized (e.g. agreement rules), or that statistical information is generated (e.g. co-occurrence of words)**

# Natural Language Processing and Web-related SLQ (2/2)

**Standard/ Guideline/ Objective**

- Adhere to standard grammar

**Linguistic Phenomenon**

- Agreement in number (determiner and noun)

**Formalized linguistic Knowledge**

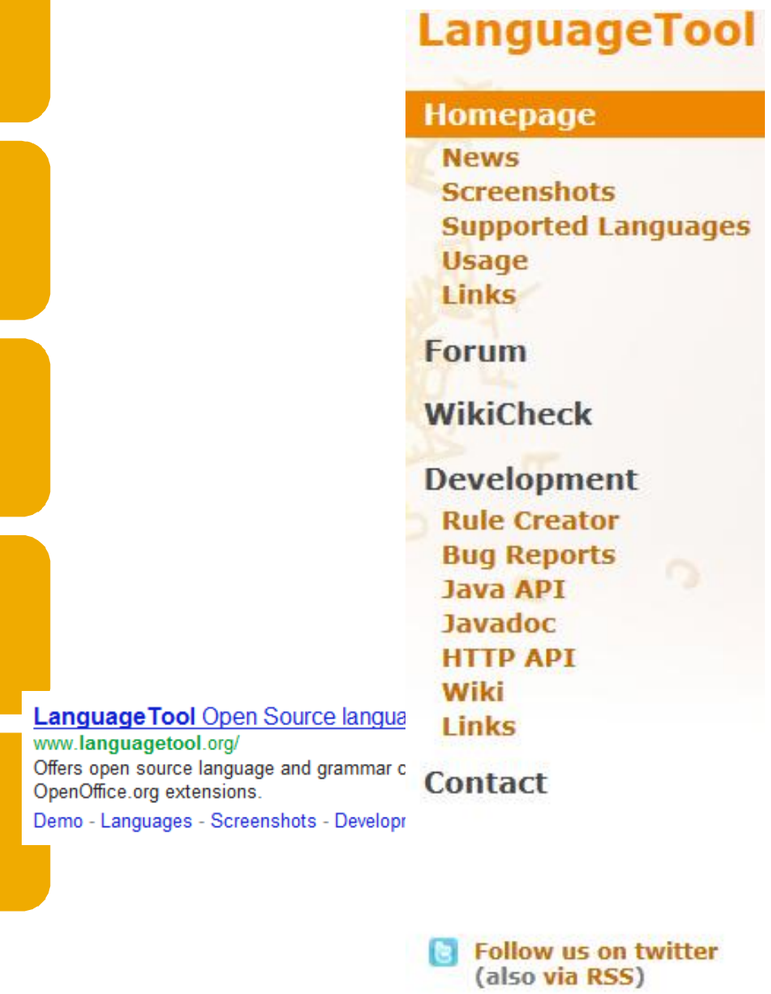
```
<unify>
    <feature id="gender">
        <type id="masc"/>
    </feature>
    <feature id="number">
        <type id="singular"/>
    </feature>
   <token/>
   <token>foo</token>
</unify>
```

| Area | Example |
|---|---|
| Spelling | Alway => Always |
| Terminology | Screen => View |
| Grammar | the program run => the program runs |
| Style | Avoid latin expressions (like *etc.*) |

# An Open Source tool for SLQ (1/3)

<S>типов[тип/NN:Masc:PL:R, тип/NN:Masc:PL:V]
сделок[сделка/NN:Fem:PL:R…</S>]

- Based on NLP (e.g. part-of-speech tagging)
- Rules-based (rules describe what shall be detected)
- English, French, German, Polish, Dutch, Romanian, and other languages (approx. 30)
- Implements also language-independent, and supports bi-lingual checks
- Support for draft W3C Internationalization Tag Set 2.0



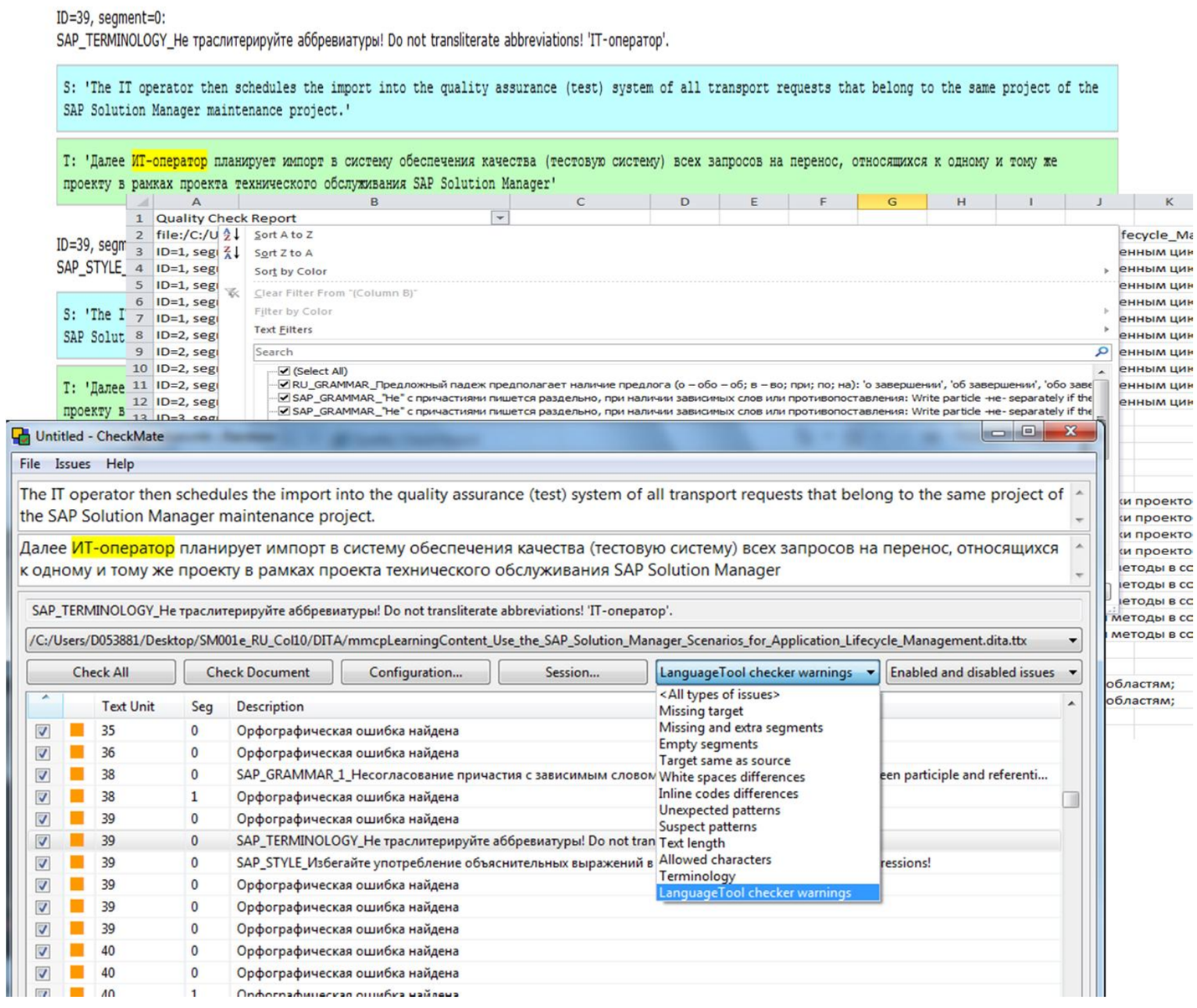
# An Open Source tool for SLQ (2/3)

- From within Host Application/Embedded (e.g. in OpenOffice/LibreOffice editor)
- Stand-alone via GUI
- Stand-alone via system tray
- Embedded as Java library
- Via output or report in XML-based format
- Coupled as HTTP-accessible service (e.g. from Okapi tools)
- Via a browser plug-in (Firefox)

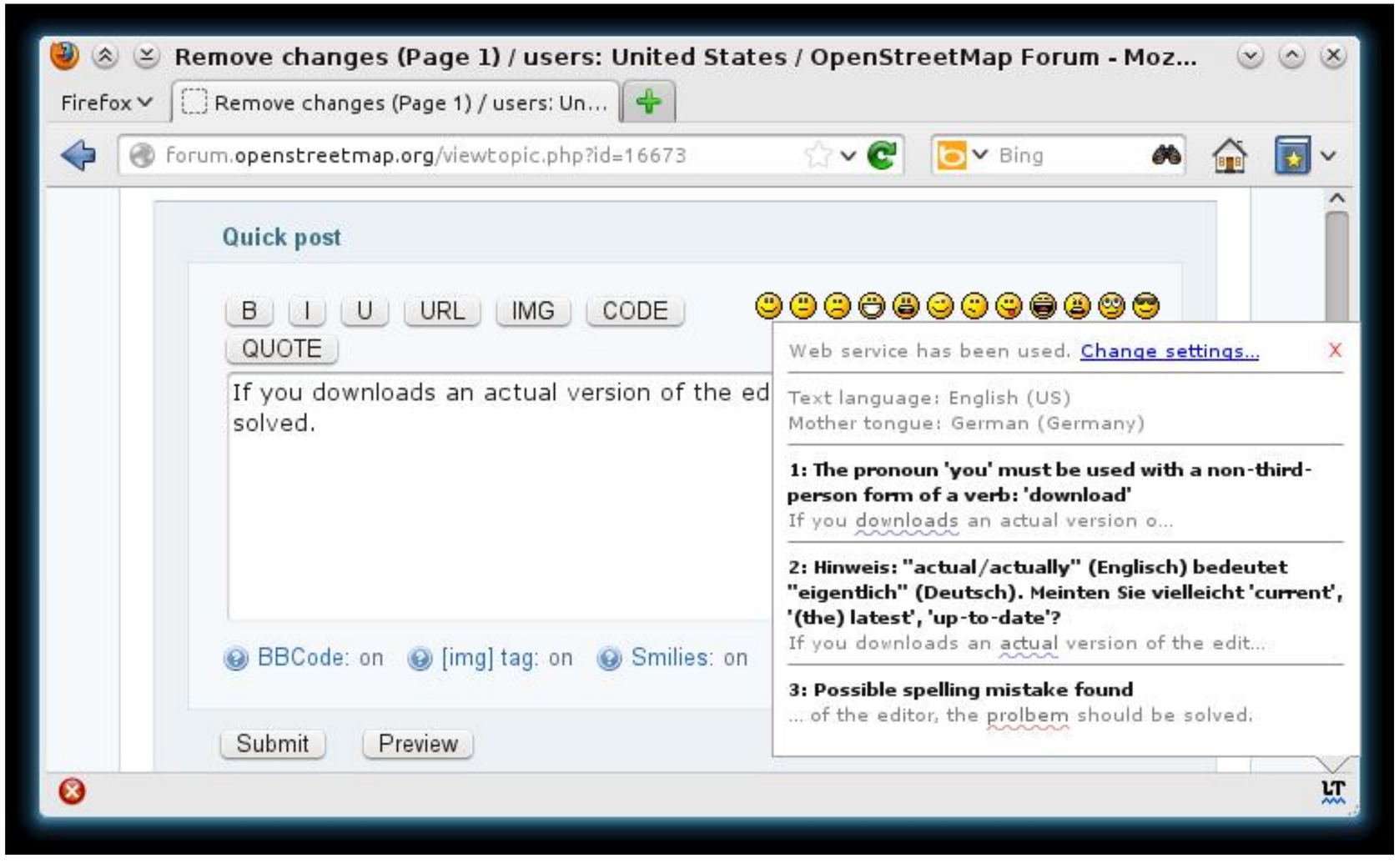
**LanguageTool Integration**

- LanguageTool for vim
- LanguageTool for LyX
- LanguageTool plugin for OmegaT source)
- LanguageTool in CheckMate used
- LanguageToolFx for Firefox
- LanguageTool for Thunderbird
- LanguageTool for Emacs

# An Open Source tool for SLQ (3/3)



# An Open Source tool for SLQ (4/4)




https://addons.mozilla.org/de/firefox/addon/languagetoolfx/

# Experiences from real-world deployments – Enterprise Scenario (1/3)

1. **в течении => в течение**

   **Ending „ии“=>“ие“;  Spelling/Orthography**

2. **выбирать календарь=> выберите**

   **Imperative mood formation, Parenthesis/Explanations; Style**

3. **текстовая документ => текстовый документ**

   **Gender agreement  (Adj. => Noun); Grammar**

4. **Например? вознаграждение... Например, вознаграждение…**

   **Comma after introductory phrases; Punctuation**

5. **Invalid: алерт, Valid: предупреждение**

   **Invalid terms, transliteration**

# Experiences from real-world deployments – Enterprise Scenario (2/3)

| | |
|---|---|
| **Easy** | • Error detection involving preposition preceding the verb |
| Example | • Для (Prep) => обеспечить (Verb)<br>• SENT_START Для[длить/DPT:Real, для/PREP] обеспечить[обеспечить/VB:INF,] |
| Cause | • Tagger Information is sufficient for the successful error detection |
| **Hard** | • Checking agreement of participle with reference is difficult for long range/non-local constructs |
| Example | • Период (<- reference noun) в минутах или часах, показывающая (<- participle) продолжительность времени… |
| Cause | • General limitation of LanguageTool – Information on syntactic constructs is not available |
| **Impossible** | • Suggestion/correction proposal involving participles not possible if singular form is required |
| Example | • Период, показывающие (<- participle)<br>• SENT_START Период[период/NN:Masc:Sin:Nom,период/NN:Masc:Sin:V],[,] показывающие[показывать/PT:Real:PL:Nom,показывать/PT:Real:PL:V,] |
| Cause | • Limitation of morphological capabilities of LanguageTool – Generation of singular form not possible |

# Experiences from real-world deployments – Enterprise Scenario (3/3)

Accuracy = recall & precision

A beter live (correct: A better life)

2 errors found = 100% recall

1 error found = 50% recall

Recall = # hits / # items

Example: 10 / 100 = 0.1 = 10%

3 errors found = 66.6% precision

Precision = # relevant hits / # hits

Example: 5 / 10 = 0.5 = 50%

Aside: Recall and precision are most often expressed as numbers between 0 and 1 – not as percentages.

| Russian | Recall | Precision |
|---|---|---|
| Orthography | n/a | n/a |
| Style | 100% | 89% |
| Grammar | 93% | 28% |
| Punctuation | 66% | 50% |
| Terminology | 67% | 92% |

# Experiences from real-world deployments – Public Service/Easy-to-Read (1/2)

14% - 33% functional analphabets – Beneficiaries of easy-to-read

```
<rule id="GENITIV-ARTIKEL">
<pattern>
  <token postag_regexp="yes"
postag="SUB:.*"/>

  <token postag_regexp="yes"
postag="ART:(DEF|IND):GEN:.*" skip="-
1"/>

  <token postag_regexp="yes"
postag="SUB:GEN:.*"/>
</pattern>

<message>Genitiv gefunden:
&quot;<match no="2"/>&quot; Vermeiden
Sie den Genitiv.</message>

</rule>
```

```
 <rule id="GENITIV-
POSSESSIVPRONOMEN">
<pattern>
  <token postag_regexp="yes"
postag="SUB:.*"/>

        <token postag_regexp="yes"
postag="PRO:POS:GEN:.*" skip="-1"/>

        <token postag_regexp="yes"
postag="SUB:GEN:.*"/>

        </pattern>

        <message>Genitiv gefunden:
&quot;<match no="2"/>&quot; Vermeiden
Sie den Genitiv.</message>

</rule>
```

Courtesy of Annika Nietzio

# Experiences from real-world deployments – Public Service/Easy-to-Read (2/2)

## Conclusions/Outlook/Contact

**Linguistic quality is scenario-dependant, and multiplies on the web**

**NLP-based automation for linguistic quality is available in the open source domain**

**The easy-to-read scenario is an important one – and needs your help**

**Let us know if you have any questions, ideas etc.**

Thank you!

christian.lieske@sap.com
inna.nickel@sap.com
naber@danielnaber.de

# Pointers

**W3C Easy-to-Read Symposium 2012 (http://www.w3.org/WAI/RD/2012/easy-to-read/#proceed)**

**How Long Is a Short Sentence? – A Linguistic Approach to Definition and Validation of Rules for Easy-to-Read Material (http://www.springerlink.com/content/t7015647p2x33380/)**

**European-dimension globale Dimension (e.g. French http://www.inclusion-europe.org/documents/100.pdf)**

**Rules, technical writing and Machine Translation (http://2011.xinnovations.de/tl_files/xinnovations.2011/slides/1909/w3c/06%20Melanie%20Siegel.pdf)**

**Scaling via Language Industry Experiences (http://2011.xinnovations.de/tl_files/xinnovations.2011/slides/1909/w3c/04%20Christian%20Lieske.pdf)**

# Abstract/Storyline

**Textual content still dominates the Web. The linguistic quality of textual content – correct spelling, terminology, grammar, style … – is of uttermost importance for various content-related processes. Linguistic quality is not universal, rather it is scenario-dependant and for example different in an enterprise scenario, than in a public service scenario. Human activities such as translation and reception as well as activities performed by software agents (e.g. search engines and Machine Translation systems) become more accurate, and cost-efficient if they operate on high-quality content. Given the volume of content on the Web, automation is important for linguistic quality management.**

**Viable automated linguistic quality management relies on so-called Natural Language Processing (NLP). Accurate NLP today requires adaptation/tailoring for the scenario at hand. With so-called rule-based/symbolic NLP this adaptation takes the shape of representing linguistic phenomena in a formalism that operates on linguistic entities such as part-of-speech tags.**

**LanguageTool is an adaptable open-source, NLP-based linguistic quality assurance tool. It offers support for approximately 30 languages, and can be used in a variety of client-server scenarios – amongst others via a browser plug-in. The body of knowledge related to adapting LanguageTool in real-world scenarios (e.g. enterprise Scenarios, and public service/easy-to-read Scenarios) is growing. LanguageTool has implemented support for the W3C Internationalization Tag Set (ITS) 2.0 that is currently under development.**

# Disclaimer

**All product and service names mentioned and associated logos displayed are the trademarks of their respective companies. Data contained in this document serves informational purposes only. National product specifications may vary.**

**This document may contain only intended strategies, developments, and is not intended to be binding upon the authors or their employers to any particular course of business, product strategy, and/or development. The authors or their employers assume no responsibility for errors or omissions in this document. The authors or their employers do not warrant the accuracy or completeness of the information, text, graphics, links, or other items contained within this material. This document is provided without a warranty of any kind, either express or implied, including but not limited to the implied warranties of merchantability, fitness for a particular purpose, or non-infringement.**

**The authors or their employers shall have no liability for damages of any kind including without limitation direct, special, indirect, or consequential damages that may result from the use of these materials. This limitation shall not apply in cases of intent or gross negligence.**

**The authors have no control over the information that you may access through the use of hot links contained in these materials and does not endorse your use of third-party Web pages nor provide any warranty whatsoever relating to third-party Web pages.**