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

Addendum for marketing

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

Guidelines for (translation into) Russian

Validated terminology

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 richtet sich an Personen, "die Entscheidungen, Informationen, Anfragen und andere Mitteilungen an Bürger vorzulegen haben".

Datenbank für bürgernahe Verwaltungssprache der Bundesbehörden

Beschäftigten der Bundesbehörden, die in der Datenbank "Verständlichkeit in der Verwaltungssprache" nutzen. Die Datenbank wurde im Rahmen des Programms "Verständlichkeit" von der Universität Bochum entwickelt.
Scenario-dependant linguistic quality and the multilingual Web

‘Link’

Is it “Read more here”,
is it “Set a link”,
is it …?
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)

<table>
<thead>
<tr>
<th>Area</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spelling</td>
<td><strong>Always</strong> =&gt; Always</td>
</tr>
<tr>
<td>Terminology</td>
<td><strong>Screen</strong> =&gt; <strong>View</strong></td>
</tr>
<tr>
<td>Grammar</td>
<td>the program run =&gt; the program runs</td>
</tr>
<tr>
<td>Style</td>
<td>Avoid latin expressions (like <em>etc.</em>)</td>
</tr>
</tbody>
</table>
An Open Source tool for SLQ (1/4)

- 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 (2/4)

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

### 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 (3/4)
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. "бирать календарь" => "берите"
   
   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)

<table>
<thead>
<tr>
<th>Easy</th>
<th>• Error detection involving preposition preceding the verb</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>• Для (Prep) =&gt; обеспечить (Verb)</td>
</tr>
<tr>
<td></td>
<td>• SENT_START Для[дайть/DPT:Real, для/PREP] обеспечить[обеспечить/VB:INF,]</td>
</tr>
<tr>
<td></td>
<td>• Tagger Information is sufficient for the successful error detection</td>
</tr>
<tr>
<td>Hard</td>
<td>• Checking agreement of participle with reference is difficult for long range/non-local constructs</td>
</tr>
<tr>
<td></td>
<td>• Период (&lt;- reference noun) в минутах или часах, показывающая (&lt;- participle) продолжительность времени…</td>
</tr>
<tr>
<td></td>
<td>• General limitation of LanguageTool – Information on syntactic constructs is not available</td>
</tr>
<tr>
<td>Impossible</td>
<td>• Suggestion/correction proposal involving participles not possible if singular form is required</td>
</tr>
<tr>
<td></td>
<td>• Период, показывающие (&lt;- participle)</td>
</tr>
<tr>
<td></td>
<td>показывающие[показывать/PT:Real:PL:Nom,показывать/PT:Real:PL:V,]</td>
</tr>
<tr>
<td></td>
<td>• Limitation of morphological capabilities of LanguageTool – Generation of singular form not possible</td>
</tr>
</tbody>
</table>
## 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%

### Precision = # relevant hits / # hits

**Example:** 5 / 10 = 0.5 = 50%

<table>
<thead>
<tr>
<th>Russian</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Orthography</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>Style</td>
<td>100%</td>
<td>89%</td>
</tr>
<tr>
<td>Grammar</td>
<td>93%</td>
<td>28%</td>
</tr>
<tr>
<td>Punctuation</td>
<td>66%</td>
<td>50%</td>
</tr>
<tr>
<td>Terminology</td>
<td>67%</td>
<td>92%</td>
</tr>
</tbody>
</table>

Aside: Recall and precision are most often expressed as numbers between 0 and 1 – not as percentages.
Experiences from real-world deployments – Public Service/Easy-to-Read (1/2)

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

<COUNT id="GENITIV-ARTIKEL" text="Genitiv gefunden: " match="2">
  Vermeiden Sie den Genitiv.
</COUNT>

<COUNT id="GENITIV-POSSESSIVPRONOMEN" text="Genitiv gefunden: " match="2">
  Vermeiden Sie den Genitiv.
</COUNT>

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)


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.