Automated monolingual term extraction through spelling check: a case study

Uta Kreimeier
Overview

1. Initial situation
2. Implemented solution
3. Advantages and disadvantages
Olympus Soft Imaging Solutions GmbH

- Headquarters: Münster, Germany
- Microscope image processing products
- Global supply unit within the Olympus group
Software localization

- English source code
- Localization into European and Asian languages
- Simship
Olympus Soft Imaging Solutions GmbH

- Highly specialized image processing software solutions
- Software users: industry and research, e.g. life sciences, materials sciences
Customization

- Software: customizable to customers' needs
- Many dialogs created at run-time
Challenges: dynamic dialogs

- Preview for static dialogs in localization tool
- Dynamic strings: no context information
- High number of specialized technical terms in string lists

<table>
<thead>
<tr>
<th>Camera Adapter</th>
<th>TV アダプタ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optional</td>
<td>任意</td>
</tr>
<tr>
<td>Shading Correction</td>
<td>シェーディング補正</td>
</tr>
<tr>
<td>Manual Magnification Calibration</td>
<td>手動倍率キャリブレーション</td>
</tr>
<tr>
<td>Camera-To-Stage Rotation</td>
<td>カメラとステージ間の回転</td>
</tr>
<tr>
<td>XY Objective Shift</td>
<td>XY 対物レンズのシフト</td>
</tr>
<tr>
<td>Exposure Correction</td>
<td>露出補正</td>
</tr>
<tr>
<td>Stage Limits</td>
<td>ステージのリミット</td>
</tr>
<tr>
<td>Parfocality</td>
<td>同焦点</td>
</tr>
<tr>
<td>Zoom Limits</td>
<td>ズームのリミット</td>
</tr>
</tbody>
</table>
Challenges: source language

- Software is developed by international and interdisciplinary team
Ideal software development world...

- Complete specifications
- No string changes in the user interface
Reality...

- No every error message is specified
- Not every requirement is known at specification stage
- Last-minute string changes due to testing feedback and additional requirements
Some resulting challenges…

- Text pool is growing daily at localization stage
- Additional new terms during localization phase
- Overlooked English language problems, e.g. spelling mistakes
Source text challenges

• Source text problems have to be logged in the bug tracking tool
• English strings are changed by the software developers only
• Language bugs are often fixed at the end
  → additional string changes
Summary: initial situation

a) Specialized terminology without context information:
   → translators’ questions, translation problems

b) Spelling mistakes
   → reporting and fixing effort, generate additional late strings changes
Overview

1. Initial situation
2. Implemented solution
3. Advantages and disadvantages
Automated term extraction with spelling check

- Commit
  - Localization tool
    - String properties
      - Technical strings
      - Translatable strings
      - Spelling check
        - Known strings
        - Unknown strings
  - Terminology database
    - Custom dictionary
    - Internal dictionary
Build process

- Daily build in all languages
- Every night, all strings in the localization tool are updated
Prerequisite: string properties

- Non-translatable text is marked accordingly
- Prerequisite for many whitebox tests
String properties: automation

- Software developers supply keywords in technical strings
- During nightly build process, string properties are set automatically in the localization tool
Spelling checker: dictionaries

- Part of every spelling checker: integrated internal dictionary
- Additional custom dictionary for company-specific words
Custom dictionary: how to maintain?

- Growing daily due to new strings
- Difficulty to maintain manually with distributed team, even with version control
Solution: Terminology database

- Custom dictionary is generated from the terminology database
- English entries with status "approved" are exported as a flat list
- English entries with status “rejected” (deprecated) are not exported
- Automated nightly export
Terminology database challenges

- Spelling checker has no complex linguistic intelligence
- Resulting challenges, e.g. plural formations or multi-word terms
- What to do with words that have to be part of the custom dictionary, but are not real terms?
Terminology database management

- Entries "spell checker only"
- Limited visibility, depends on user rights
Checking output: unknown strings

- Technical strings
- Spelling mistakes
- New terms
Error message: technical string

2012-12-05 07:06:58,753 [WARN ] - WorkflowCommonGUI: Font in Dialog 3040 111: Spelling error "Dlg" (‘MS Shell Dlg(DS_SHELLFONT)’)

→ String has to be marked accordingly in localization tool

• Software developers have to set correct keyword for technical strings
• In some cases, software developers have to contact software localization so that correct string properties are set manually
Error message: spelling mistake

2012-12-02 07:07:01,171 [WARN ] - dotSlide: Static 5902 in Dialog 3058 111: Spelling error "Mininimum" ('Mininimum size for separate scan areas per sub-sample:\')

→ Spelling mistake has to be corrected

• Software developers changes string in the source code
• If necessary, software localization provides advice
Error message: new word

2012-12-06 07:07:22,113 [WARN ] - WordAutomation: String
"PlaceholderSize_0_1_Picas" in StringTable.NET
"WordAutomation.Properties.Resources" 111: Spelling error "pi" ('Placeholder size: {0} pi x {1} pi')

→ Term has to be added to the terminology database

• Software developers have to provide context information
• Software localization creates new entry in terminology database
Terminology database

- Displays definitions and additional information
- Helps translators to establish correct target language equivalents
- Used as dictionary by in-house staff
Overview

1. Initial situation
2. Implemented solution
3. Advantages and disadvantages
Limitations

- **Compounds**
  If all parts of the compound are known, the compound term will not be recognized

- **Homonyms**
  one designation for more than one concept
  Example: well (adverb/ noun)

- **Limitations of the spell checker technology:** plural formation, full stop

- **No linguistic intelligence**
Not all new terms are found…

→ Additional manual terminology extraction

Proactive – prior to localization
• New core marketing terms are communicated early
• Visual check of new dialogs

Reactive – during/ after localization
• Through translators‘ questions (query management)
Automated monolingual term extraction through spelling check: a case study

- Realization with existing tools
- Responsibility is shifted upstream
- Immediate feedback for software developers
- Terms are added to the terminology database as soon as they appear
- Context information is delivered automatically through a clear process
- Clear workflows
- Reduction of workload in final phases of project
- Reduction of manual work
Thank you

Uta Kreimeier
Translation Coordinator

Olympus Soft Imaging Solutions GmbH
Johann-Krane-Weg 39
48149 Münster, Germany

uta.kreimeier@olympus-sis.com