Terminology Management in Small and Medium-sized LSPs

Andraž Repar
Quality Manager
Iolar, Ljubljana, Slovenia
Iolar

- Headquartered in Ljubljana, Slovenia
- On the market since 1992
- 5 offices, 40 full-time employees
- Focus on life sciences, automotive industry, engineering, public sector
- Member of Gala, Elia, Translators without Borders
- Organizer of the TTT Conference (9-10 November in Bled, Slovenia)
About the speaker

- Studied translation at the University of Ljubljana (English, German)
- Started as translator/reviser at Iolar
- Later began working as Quality Manager at Iolar
- ECQA Terminology Manager
- Doing PhD studies at the International Postgraduate School at the Jozef Stefan Institute in Ljubljana, Slovenia
- Focus on natural language processing and machine learning - with practical application on terminology extraction
Outline

● The role of terminology in the translation industry
● The current state of terminology solutions
● Terminology processes at Iolar
● An ideal terminology solution for LSPs
● Iolar’s attempts at solving the terminology puzzle
Why terminology?

- Digital transformation in the translation industry
  - Translation memories
  - QA checkers, spell-checkers
  - Project management automation
  - Machine translation
  - Terminology?

Neural Machine Translation—A Practical Report from the Front Line, John Tinsley, Iconic CEO
Why terminology?

- SDL Translation Technology Insights Series: Quality
  - 2784 respondents in 115 countries
  - Freelance translators, LSPs, Corporate buyers

Terminology is the top challenge

We asked those who face rework to choose the three most common reasons for it.

*Inconsistencies in the use of terminology* emerged as the top cause, chosen by 48%.

The results also show that getting terminology right is more of an issue for corporates than the other two groups, with 60% citing it among their top three reasons for rework.

Prioritize terminology management

56% say that terminology management is a priority.

55% say that terminology management tools seem too complicated or time-consuming.
Terminology for LSPs

- Two trends
  - The number of projects is increasing
  - Deadline are getting shorter

- Consequence ➔ less time for:
  - Style guides
  - Glossaries
  - Terminology management
Terminology for LSPs

What are the specific aspects of terminology work at Iolar?

- Sub-contracting for LSPs vs. direct customers
- Different CAT tools
- No available termlist
- Short deadlines
- Multiple translators working simultaneously
- Customer involvement

What is the best tool for all this?
## Terminology for LSPs

<table>
<thead>
<tr>
<th>LSP</th>
<th>Corporate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple domains</td>
<td>Limited set of domains</td>
</tr>
<tr>
<td>Various file formats</td>
<td>Predictable file formats</td>
</tr>
<tr>
<td>Limited time-frame</td>
<td>Long-term terminology management</td>
</tr>
<tr>
<td>Limited access to SMEs</td>
<td>In-house SMEs</td>
</tr>
<tr>
<td>Unlimited languages</td>
<td>Limited set of languages</td>
</tr>
<tr>
<td>(Often) no dedicated terminology manager</td>
<td>(At least one) dedicated terminology manager</td>
</tr>
<tr>
<td>“loose” term definition</td>
<td>“strict” term definition</td>
</tr>
</tbody>
</table>
Terminology for LSPs

What does Iolar use for terminology work?

- SDL Multiterm
- SDL Appstore apps
  - Glossary Converter
  - Glossary Plugin
  - TermExcelerator
- ApSIC Xbench
- Microsoft Excel/Google Sheets
## Terminology for LSPs

Multiple translators working together

<table>
<thead>
<tr>
<th>Filename</th>
<th>SL</th>
<th>EN</th>
<th>Comment</th>
<th>Confirmation</th>
<th>Client Approved</th>
<th>Translator</th>
</tr>
</thead>
<tbody>
<tr>
<td>RRP 4: Trdne dispersije</td>
<td>zdravilo</td>
<td>medicinal product / medicine idrug</td>
<td></td>
<td></td>
<td></td>
<td>Aleksandra</td>
</tr>
<tr>
<td>RRP 2: Scale up procesov za načrtovanje delcev zdravilnih učinkovin (TRL 5-6)</td>
<td>korenno zdravilo</td>
<td>finished product</td>
<td></td>
<td></td>
<td></td>
<td>Aleksandra</td>
</tr>
<tr>
<td>načrtovanie delcev</td>
<td>particle design</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Tea</td>
</tr>
</tbody>
</table>
Terminology for LSPs

Google Sheets

- Positives
  - Familiar interface
  - Easy sharing
  - Simultaneous editing
  - Nice commenting system
  - Chat support

- (Potential) negatives
  - Cloud-based
  - It’s Google
Terminology for LSPs

But still...

- How do you manage terminology used in different CAT-tools?
- How to ensure company-wide access to terminology resources?
Terminology for LSPs

An ideal terminology system for Iolar

- Integrates with (almost) all CAT-tools
- On-the-fly editing (from any CAT-tool)
- Familiar collaboration workflow
- Terminology extraction
- Extraction of definitions, key phrases, examples
- Domain recognition and termbase suggestions

Part of a larger translation workflow system?
In 2015, Iolar decided to focus on terminology and started developing its own terminology management and extraction system in cooperation with the JSI Institute:

- Support for “our” languages
- Terminology database
- (Bilingual) terminology extraction
- Good example extraction
- Definition extraction
- Domain extraction

![TermIolar](image)
TermIolar

- Web application
- Single database
- TBX and Multiterm XML compliant
- Import/export functionality
- Extraction module
- Broad file format support
- API support
TermIolar

- Terminology management
- (Bilingual) terminology extraction
- Definition extraction
- Domain extraction
- Good example extraction
Terminology extraction

- **Statistical**: count frequent words/phrases
- **Linguistic**: find specific patterns (e.g. Adj + Noun)
- **Hybrid**: find patterns and count their frequency

What is a term?

- ISO 1087: a verbal designation of a general concept in a specific subject field
- Unithood vs. termhood
  - A new supervised machine learning algorithm
  - Supervised machine learning algorithm
  - Machine learning algorithm
  - Supervised machine learning
  - Machine learning
  - Learning algorithm?
Termhood

- A term is a verbal designation of a general concept in a specific subject field
- A term is used *more often* in its specific subject field than in the general language
- Compare the frequency of a term in a specific subject field corpus and its frequency in a general language corpus

\[ W(a) = \frac{f_a}{n} \times \sum_{1}^{n} \left( \log \frac{f_{n,D}}{N_D} - \log \frac{f_{n,R}}{N_R} \right) \]

Bilingual terminology extraction

\( n \) is the number of tokens in a term
\( f_a \) is the absolute frequency of the candidate term in the domain-specific corpus
\( f_{n,D} \) and \( f_{n,R} \) are the frequencies of each constituent word in the domain-specific and the general language reference corpus
\( N_D \) and \( N_R \) are the sizes of these two corpora in tokens

Vintar, 2010
Bilingual term alignment

Two lists of terms and a parallel corpus

- Co-frequency
- logDice
- Mutual information
- Bag-of-equivalents
- MT phrase table
- Bilingual bag-of-words
- Binary classification

<table>
<thead>
<tr>
<th>L1 term</th>
<th>L2 term</th>
<th>Co-freq</th>
<th>L1 freq</th>
<th>L2 freq</th>
<th>MI</th>
<th>dice</th>
</tr>
</thead>
<tbody>
<tr>
<td>share</td>
<td>delnica</td>
<td>2305</td>
<td>3564</td>
<td>4946</td>
<td>-12,90077711</td>
<td>0,541715629</td>
</tr>
<tr>
<td>market</td>
<td>trg</td>
<td>2114</td>
<td>3105</td>
<td>3148</td>
<td>-12,17483772</td>
<td>0,676155445</td>
</tr>
<tr>
<td>asset</td>
<td>sredstvo</td>
<td>1894</td>
<td>2443</td>
<td>2562</td>
<td>-11,69027229</td>
<td>0,756843157</td>
</tr>
<tr>
<td>fund</td>
<td>sklad</td>
<td>1447</td>
<td>2355</td>
<td>4290</td>
<td>-12,76942407</td>
<td>0,435515425</td>
</tr>
<tr>
<td>share</td>
<td>razred</td>
<td>1260</td>
<td>3564</td>
<td>2122</td>
<td>-12,55128254</td>
<td>0,443193809</td>
</tr>
</tbody>
</table>
Term extraction - Results

- Voting: 7 methods - 7 votes
- Translation memory
  - 18,144 units
  - 352,868 words
- 1838 extracted term pairs
- 330 of them with at least 4 votes
- Precision of 71%
  - Both terms have to be correct
  - Term alignment has to be correct
- Some incorrect alignments are easy to fix!

<table>
<thead>
<tr>
<th>EN term</th>
<th>SL term</th>
<th>#VOTES</th>
<th>MANUAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>volatility</td>
<td>volatilnost</td>
<td>7</td>
<td>TRUE</td>
</tr>
<tr>
<td>valuation</td>
<td>vrednotenje</td>
<td>5</td>
<td>TRUE</td>
</tr>
<tr>
<td>utility</td>
<td>javen storitev</td>
<td>4</td>
<td>TRUE</td>
</tr>
<tr>
<td>usd</td>
<td>usd</td>
<td>4</td>
<td>TRUE</td>
</tr>
<tr>
<td>undertaking for collective investment</td>
<td>kolektivni naložbeni podjem</td>
<td>7</td>
<td>TRUE</td>
</tr>
<tr>
<td>underlying asset</td>
<td>osnovno sredство</td>
<td>5</td>
<td>TRUE</td>
</tr>
<tr>
<td>trend</td>
<td>trend</td>
<td>6</td>
<td>TRUE</td>
</tr>
<tr>
<td>transfer</td>
<td>prenos</td>
<td>6</td>
<td>TRUE</td>
</tr>
<tr>
<td>transaction costs</td>
<td>strošek transakcija</td>
<td>4</td>
<td>TRUE</td>
</tr>
<tr>
<td>transaction</td>
<td>transakcija</td>
<td>6</td>
<td>TRUE</td>
</tr>
<tr>
<td>technology</td>
<td>tehnologija</td>
<td>5</td>
<td>TRUE</td>
</tr>
<tr>
<td>target funds</td>
<td>ciljni sklad</td>
<td>4</td>
<td>TRUE</td>
</tr>
<tr>
<td>swap</td>
<td>zamenjava</td>
<td>6</td>
<td>TRUE</td>
</tr>
<tr>
<td>supervisory authority</td>
<td>nadzorni organ</td>
<td>7</td>
<td>TRUE</td>
</tr>
<tr>
<td>subscription price</td>
<td>vpisna cena</td>
<td>7</td>
<td>TRUE</td>
</tr>
<tr>
<td>subscription</td>
<td>vpis</td>
<td>6</td>
<td>TRUE</td>
</tr>
<tr>
<td>subfund</td>
<td>podsklad</td>
<td>4</td>
<td>TRUE</td>
</tr>
<tr>
<td>strategy</td>
<td>strategija</td>
<td>4</td>
<td>TRUE</td>
</tr>
<tr>
<td>stock selection</td>
<td>izbira delnic</td>
<td>4</td>
<td>TRUE</td>
</tr>
<tr>
<td>stock exchange</td>
<td>borza</td>
<td>4</td>
<td>TRUE</td>
</tr>
<tr>
<td>sovereign debt crisis</td>
<td>kriza državnih dolgov</td>
<td>7</td>
<td>TRUE</td>
</tr>
<tr>
<td>simplified prospectus</td>
<td>izvleček prospekta</td>
<td>5</td>
<td>TRUE</td>
</tr>
</tbody>
</table>
TermIolar

- Terminology management
- (Bilingual) terminology extraction
- Definition extraction
- Domain extraction
- Good example extraction
Good example extraction

What is a good example?

Only 1 in 4 segments can be considered a good example!

According to Kilgarriff et al. (2008), a good dictionary example must be:

- typical, exhibiting frequent and well-dispersed patterns of usage
- informative, helping to elucidate the definition
- intelligible to learners, avoiding gratuitously difficult lexis and structures, puzzling or distracting names, anaphoric references or other deictics which cannot be understood without access to the wider context
Good example extraction

- Machine learning approach - binary classification
- Term extraction
- Manual annotation (1332 bilingual segments, 962 negative, 370 positive)
- (Mostly) language independent features (segment length, term position, number of commas...)
- Several ML algorithms (Naïve Bayes, decision trees, k-nearest neighbours, support vector machine)

Result: significant improvement in positive class precision → 4 out 5 positive examples classified correctly
Conclusion

- Terminology work is slow, difficult and expensive
- Existing terminology solutions cannot fulfill the needs of LSPs
- Terminology work involves a multitude of different apps
- Terminology activities should be integrated into regular translation workflows

Questions?
Your opinion is important to us! Please tell us what you thought of the lecture. We look forward to your feedback via smartphone or tablet under

http://tekom08.honestly.de

or scan the QR code

The feedback tool will be available even after the conference!