Terminology Work — Breaking the Barriers
Automatic Term Extraction

An Stuyven – Skrivanek Group
Vancouver, October 29, 2014
Experience with Terminology

Skrivanek has been doing for 20 years

- terminology work with small and large clients
- covering the whole terminology workflow:

  - Terminology Process Consultancy
  - Clearing the objectives and conditions (existing terminology, type and size of the content, quality requirements, tools, integration, cost, …)
  - Input and validation of existing glossaries
  - Terminology extraction and translation
  - Online terminology sharing (tools) with whole corporation
  - Terminology maintenance and ongoing processing
Typical Barriers

- Input
- Varying candidate quality
- Output
- Compatibility
- User-friendliness of sharing method
Term Creation and Validation

1. Define and verify process
2. Identify and create corpus
3. Extract term candidates
4. Definition tasks
5. Termbase
6. Cleanup and validate terms
7. Translate Terms
Term extraction

- Manual extraction

- Automatic extraction: tools
e.g.: TerMine,
+ Most TMS tools have this function (SDL Trados, MemoQ,...)

- Main problem:
  high percentage of „noise and silence“
Term Extraction

Statistical Approach

Most common
Examples: TerMine, Fivefilters Term Extraction, SDL Multiterm Extract, …

Based on frequency
- Language independent
- Issues
  - The frequency threshold must be specified
  - Frequency does not necessarily means importance
  - Much „noise“ and „silence“ – extensive manual postprocessing

Linguistic Approach

Based on rules and dictionaries

- Not available for all languages
- Issues
  - Loans
  - Synonyms, variants, abbreviations
  - Ellipses
  - Improper usage
Term extraction with qTerm (MemoQ)
**Term extraction with qTerm - Candidates**

<table>
<thead>
<tr>
<th>$</th>
<th>#</th>
<th>Status</th>
<th>Source</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>18.55</td>
<td>63</td>
<td>Candidate</td>
<td>accident</td>
<td>Unfallspezialist</td>
</tr>
<tr>
<td>3.33</td>
<td>4</td>
<td>Candidate</td>
<td>accident Business</td>
<td></td>
</tr>
<tr>
<td>5.00</td>
<td>6</td>
<td>Candidate</td>
<td>accident data</td>
<td></td>
</tr>
<tr>
<td>9.18</td>
<td>6</td>
<td>Candidate</td>
<td>accident data collection</td>
<td></td>
</tr>
<tr>
<td>14.13</td>
<td>6</td>
<td>Candidate</td>
<td>accident data collection form</td>
<td></td>
</tr>
<tr>
<td>10.03</td>
<td>13</td>
<td>Accepted</td>
<td>Accident specialist</td>
<td></td>
</tr>
<tr>
<td>7.50</td>
<td>9</td>
<td>Candidate</td>
<td>accident specialists</td>
<td></td>
</tr>
<tr>
<td>9.72</td>
<td>33</td>
<td>Candidate</td>
<td>Accidental</td>
<td></td>
</tr>
<tr>
<td>27.48</td>
<td>33</td>
<td>Candidate</td>
<td>Accidental Damage</td>
<td></td>
</tr>
</tbody>
</table>

The employees from the department can participate on the involvement of an accident specialist (preferably a multi-brand partner with ŠKODA in service). Können Mitarbeiter aus dem Fachbereich bei einem Einsatz eines Unfallspezialisten (am Besten ein Mehrmarkenpartner mit ŠKODA im Service) teilnehmen.

**Professionalization of ADM without**

Source: SMB_EN.pptx (TranslationDoc)
Term extraction with SDL Multiterm (Trados)
The Sketch Engine is for anyone wanting to research how words behave. It is a Corpus Query System.
Term Finding with Sketchengine

Linguistic approach
Tokenising, lemmatising, POS-tagging

Statistical approach
Frequency in domain corpus vs reference corpus

- Around 70 languages
- Company is specialized in building these reference corpora
Term Finding with Sketchengine

Term candidates for a domain, in a language, can be found by:

- Taking a corpus for the domain and a reference corpus for the language
- Identifying the grammatical shape of a term in the language
- Tokenising, lemmatising and POS-tagging both corpora
- Identifying and counting the items in each corpus which match the grammatical shape
- For each item in the domain corpus, comparing its frequency with its frequency in the reference corpus

-> Items with highest domain:reference ratio are the top term candidates
### Corpora

<table>
<thead>
<tr>
<th>Language</th>
<th>Corpus name</th>
<th>Tokens</th>
<th>Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chinese, Simplified</td>
<td>zhTenTen11</td>
<td>2,106,661,021</td>
<td>1,729,867,455</td>
</tr>
<tr>
<td>English</td>
<td>British National Corpus</td>
<td>112,181,015</td>
<td>96,048,950</td>
</tr>
<tr>
<td>Japanese</td>
<td>jpTenTen11 [SUW]</td>
<td>10,321,875,664</td>
<td>8,432,256,386</td>
</tr>
<tr>
<td>Spanish</td>
<td>esTenTen11 (European, Freeling)</td>
<td>2,341,159,406</td>
<td>1,991,879,282</td>
</tr>
</tbody>
</table>

### My corpora

<table>
<thead>
<tr>
<th>Language</th>
<th>Corpus name</th>
<th>Configuration template</th>
<th>Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>EN-DE TB Extraction</td>
<td>TreeTagger for English</td>
<td>0</td>
</tr>
</tbody>
</table>

[Create corpus] [WebBootCaT]
<table>
<thead>
<tr>
<th>Keywords</th>
<th>Score</th>
<th>F</th>
<th>RefF</th>
<th>Terms</th>
<th>Score</th>
<th>F</th>
<th>RefF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Skoda</td>
<td>9,628.64</td>
<td>27</td>
<td>860</td>
<td>accidental damage</td>
<td>3,006.55</td>
<td>10</td>
<td>3,443</td>
</tr>
<tr>
<td>adm</td>
<td>3,359.92</td>
<td>13</td>
<td>6,114</td>
<td>accident specialist</td>
<td>2,644.09</td>
<td>7</td>
<td>94</td>
</tr>
<tr>
<td>autorechtsaktuell</td>
<td>1,521.91</td>
<td>4</td>
<td>0</td>
<td>collection form</td>
<td>1,874.03</td>
<td>5</td>
<td>206</td>
</tr>
<tr>
<td>ohne</td>
<td>1,322.00</td>
<td>5</td>
<td>5,691</td>
<td>data collection form</td>
<td>1,518.87</td>
<td>4</td>
<td>32</td>
</tr>
<tr>
<td>accidental</td>
<td>1,276.94</td>
<td>23</td>
<td>75,857</td>
<td>authorization agreement</td>
<td>1,512.83</td>
<td>4</td>
<td>89</td>
</tr>
<tr>
<td>usm</td>
<td>1,043.86</td>
<td>4</td>
<td>5,939</td>
<td>damage management</td>
<td>1,132.62</td>
<td>3</td>
<td>115</td>
</tr>
<tr>
<td>aggrieved</td>
<td>887.59</td>
<td>4</td>
<td>9,268</td>
<td>claim settlement</td>
<td>1,086.28</td>
<td>3</td>
<td>672</td>
</tr>
<tr>
<td>constructio</td>
<td>750.69</td>
<td>2</td>
<td>186</td>
<td>work accident</td>
<td>1,059.08</td>
<td>3</td>
<td>1,022</td>
</tr>
<tr>
<td>cooperations</td>
<td>716.64</td>
<td>2</td>
<td>811</td>
<td>external service</td>
<td>1,058.09</td>
<td>3</td>
<td>1,030</td>
</tr>
<tr>
<td>bods</td>
<td>710.81</td>
<td>2</td>
<td>924</td>
<td>data collection</td>
<td>943.53</td>
<td>4</td>
<td>7,961</td>
</tr>
<tr>
<td>cession</td>
<td>674.26</td>
<td>2</td>
<td>1,677</td>
<td>quota accident</td>
<td>761.46</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>wolfsburg</td>
<td>672.61</td>
<td>2</td>
<td>1,713</td>
<td>fax authorization agreement</td>
<td>761.46</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>dealership</td>
<td>670.26</td>
<td>11</td>
<td>67,975</td>
<td>collection form assignment</td>
<td>761.46</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>seeger</td>
<td>572.91</td>
<td>2</td>
<td>4,268</td>
<td>data collection form assignment</td>
<td>761.46</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>plausibility</td>
<td>555.63</td>
<td>2</td>
<td>4,804</td>
<td>form assignment</td>
<td>761.46</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>trainings</td>
<td>526.15</td>
<td>5</td>
<td>33,915</td>
<td>procedure audit</td>
<td>761.46</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>jens</td>
<td>488.31</td>
<td>2</td>
<td>7,254</td>
<td>residual value plausibility check</td>
<td>761.46</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>accident</td>
<td>482.83</td>
<td>59</td>
<td>589,597</td>
<td>value plausibility check</td>
<td>761.46</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>incl</td>
<td>460.66</td>
<td>2</td>
<td>8,468</td>
<td>accident data collection</td>
<td>761.46</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>
SketchEngine
Building a domain corpus

WebBootCaT

If no domain corpus is available, it can be created:

• Send „seed terms“ to a commercial search engine
• Gather the indexed pages
• Cleaning, deplicating and indexing as a corpus
• Corpus can be used for translators to find concordance examples
Rule-Based MT for Term Extraction

• Rule based Machine Translation Systems can be used for Term Extraction*

• No direct function
  Use the reverse: „unknown words“ list

• Result depends on specialization of your topic and on MT feed

As in the Paper by Uwe Mögge „Automatische Terminologieextraction, January 2007 http://works.bepress.com/cgi/viewcontent.cgi?article=1001&context=uwe_muegge
Term Extraction

- Tool choice according to requirements (bilingual extraction,

- play with settings
  - input (corpus, TM, ...)
  - frequency
  - length of terms
  - stop word lists
  - ...
Thank you for your attention!

An Stuyven
Skrivanek Group

an.stuyven@skrivanek.com