Bilingual Terminology Extraction from TMX
A State-of-the-Art Overview

Chelo Vargas-Sierra, PhD
University of Alicante, Spain
INDEX

Main points of this presentation

Key words
Overview of terms involved in the process

Evaluation
BATE under evaluation
Measures for accuracy
Quality in use model and tasks

1st point
Terminology and extractors
Terminology management
Its timeline
BATE (approaches, state of the art)

2nd point

3rd point

4th point
Results
Precision & Recall
Parameters & Questionnaire
KEY WORDS
Terms involved in the process

Parallel corpus
TMX

Alignment levels
Paragraph, sentence and word level

ATE & BATE

Precision/Recall
Getting only terms and all terms

Gold standard
Exhaustive, manually created bilingual glossary

Validation
* Term validation facility
* Which TCs are real terms?

Usability
Software used to achieve user's objectives with effectiveness, efficiency, and satisfaction

Quality in use model
ISO standard
2. Terminology & Extractors
IMPORTANCE OF TERMINOLOGY

Translators were the first professionals to be aware of term-related issues

Identify and interpret the terminology in the source text adequately

Find and use proper documentation and information resources

Retrieve and store terminological data

Identify

Find

Retrieve
TERMINOLOGY MANAGEMENT

In specialized translation

+40% Time spent to solve terminological problems (Arntz 1993, Walker 1993).
TERMINOLOGY MANAGEMENT

In specialized translation

+40% Time spent to solve terminological problems (Arntz 1993, Walker 1993).

Managing terminology (extracting, validating, importing, adding, editing, deleting, revising, updating, exporting, publishing) is a time-consuming process.
TERMINOLOGY MANAGEMENT

In specialized translation

+40% Time spent to solve terminological problems (Arntz 1993, Walker 1993).

Managing terminology (extracting, validating, importing, adding, editing, deleting, revising, updating, exporting, publishing) is a time-consuming process.

Terminology work is “on backstage”, and customer or employers may not be fully aware of their benefits for QA.
TERMINOLOGY MANAGEMENT

In specialized translation

+40% Time spent to solve terminological problems (Arntz 1993, Walker 1993).

Managing terminology (extracting, validating, importing, adding, editing, deleting, revising, updating, exporting, publishing) is a time-consuming process.

Terminology work is “on backstage”, and customer or employers may not be fully aware of their benefits for QA.

90% Return on Investment (ROI) on terminology management reported by some corporate studies (Childress, 2007; Popiolek, 2015)
TERMINOLOGY MANAGEMENT

General model por project terminology creation (Popiolek, 2015: 351)

- **Extraction**
  - List of terms **extracted** from ST
  - List of terms to **validate** (accept or reject)

- **Translation**
  - List is added to a **termbase**
  - List is **translated** and additional data added

- **Approval**
  - List approved by a person in charge of terminology
  - When the client has requested there is an additional step for client approval

Monolingual extraction & validation
Importing & looking for equivalents

General model por project terminology creation (Popiolek, 2015: 351)
TIMELINE in Terminology Management

**Preparation: TMX import**
Preparing the files and import them into the BATE

**Bilingual extraction**
List of candidate term pairs extracted from TMX
Validation (& data entry)
- List of pair of terms to validate (accept or reject terms and suggested equivalents)
- Term by term and additional data are added to a term base (Synchroterm)

Export/Import
- Export bilingual terms and additional data in an available file format (.xls, .txt, .TBX, …)
- Import output file to a TDB system (to be integrated into a MT System)
Approval
Person in charge of terminology or client

Finish
Ready to use

Ready to use
Bilingual Automatic Term Extractors

Two approaches (Foo, 2012)

**EXTRACT-ALIGN**

1st step: monolingual terminology extraction in both languages.

2nd step: cross-linguistic matching using word-alignment or co-occurrence statistics to find equivalents.

Commercial systems in this approach
Bilingual Automatic Term Extractors

Two approaches (Foo, 2012)

ALIGN-FILTER

1\textsuperscript{st} step: word-alignment on the parallel texts.

2nd step: rank the aligned units to finally select the most likely pair of candidates (statistics)

TExSIS (Macken et al, 2013)
Bilingual Automatic Term Extractors  

Academic / In-house

**90s**
- English-French **TERMIGHT** (Dagan & Church, 1994)
- English-French (Kupiek, 1993)
- English-Dutch (Eijk, 1993)
- English-French (Gaussier, 1995)
- English and Swedish (Ahrenberg et al., 1998)

**2000-2009**
- French-German (Blank, 2000)
- Japanese-English, **MNH** (Nakagawa & Mori, 2003)
- Spanish-Basque, **Elexbi** (Hernaiz et al., 2006), from a TMX;
- Spanish-German, **Autoterm** (Haller, 2008);
- English-Spanish, **Mutual Bilingual Term Extractor** (Ha et al, 2008)
- French-English, French-Italian and French-Dutch (Lefever et al., 2009)

**2010 -2016**
- Slovene and English, **Luiz** (Vintar, 2010);
- English and Swedish **ITools suite** (Foo & Merkel, 2010)
- English and German (Gojun et al., 2012).
- English, French, German, Spanish, **TTC TermSuite** (Daille, 2012)
- English-Spanish **TBXTools** (Oliver & Vázquez, 2015) (under development)
- Chinese, Czech, Dutch, English, French, German, Italian, Japanese, Korean, Polish, Portuguese, Russian, Spanish: **Sketch Engine** (Baisa et al 2015, Koval et al 2016)
Bilingual Automatic Term Extractors

Other BATE (free / comercial)

**MONOLINGUAL ATE**

- **TermExtractor** (Shimohata et al 2001)
- MemoQ’s built-in term extractor
- Déjà Vu - Lexicon
- TermoStat Web: [http://termostat.ling.umontreal.ca/](http://termostat.ling.umontreal.ca/)
- Yate (IULA)
- Okapi
- TerMine: [http://www.nactem.ac.uk/software/termine/](http://www.nactem.ac.uk/software/termine/)
- TerminologyExtractor: [https://goo.gl/yA2Cuf](https://goo.gl/yA2Cuf)
- PRoMT
- FiveFilters (web-based): [http://fivefilters.org/term-extraction/](http://fivefilters.org/term-extraction/)
- **Concordace programs**: WordSmith Tools, AntConc (free), ...

**BILINGUAL**

- Xerox Terminology Suite (2001)
- SDL Multiterm Extract
- Synchroterm
- CrossMining (Across)
- MultiTrans Term Extractor
- Similis™ (by Lingua et Machina™)
- Anchovy (by Swordfish)
- Araya Term Extractor

**Analysis software**: Sketch Engine (terminology extraction from TMX)
3. Evaluation
# MAIN FEATURES

<table>
<thead>
<tr>
<th>Feature</th>
<th>Multiterm Extract</th>
<th>SynchroTerm</th>
<th>Similis</th>
<th>SkE</th>
<th>Araya</th>
</tr>
</thead>
<tbody>
<tr>
<td>Import TMX</td>
<td>✔</td>
<td>✔</td>
<td>Trados TMX</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Extraction config.</td>
<td>✔</td>
<td>✔</td>
<td>X</td>
<td>X</td>
<td>✔</td>
</tr>
<tr>
<td>Extraction scores</td>
<td>✔</td>
<td>✔</td>
<td>X</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Validation facility</td>
<td>✔</td>
<td>✔</td>
<td>+</td>
<td>X</td>
<td>+</td>
</tr>
<tr>
<td>Term base indexation</td>
<td>✔</td>
<td>✔</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Export to TBX (xls, txt…)</td>
<td>✔</td>
<td>🌟 Others</td>
<td>🌟 Others</td>
<td>✔</td>
<td>✔</td>
</tr>
</tbody>
</table>
MEASURES FOR ACCURACY

\[
\text{PRECISION} = \frac{A}{A+C}
\]

\[
\text{RECALL} = \frac{A}{A+B}
\]
QUALITY IN USE MODEL

Characteristics (ISO-IEC 25010: 2011)

- **Satisfaction**: degree to which user needs are satisfied when a software is used in a specified context of use.
- **Efficiency**: resources expended in relation to the accuracy and completeness.
- **Effectiveness**: accuracy and completeness with which user achieves objectives.
- **Freedom from risk**: no risk for the security of users, software, context or the environment.
- **Context coverage**: degree to which the product understands the complete context of its usage. Flexibility.
6 TASKS TO EVALUATE
when performing bilingual extraction

- **CONFIGURATION**
  Setting up the extraction project

- **TMX IMPORT**
  Importing the source file

- **EXTRACTION**
  Performing the extraction to get a bilingual list

- **VALIDATION**
  Selecting the real terms.

- **RECORD CREATION**
  Creating and managing term entries

- **EXPORTATION**
  Exporting the final result for later use in CAT Systems
4. Results
## PRECISION & RECALL IN %

### Bar Chart:
- **X-axis**: Precision and Recall (%)
- **Y-axis**: Values range from 0.00 to 70.00
- **Legend**: Sketch, MTE, Synchr, Similis

### Table:
<table>
<thead>
<tr>
<th></th>
<th>EXTRACTED</th>
<th>NON-EXTRACTED</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TERMS</td>
<td>NO TERMS</td>
</tr>
<tr>
<td>Sketch</td>
<td>A 283</td>
<td>C 717</td>
</tr>
<tr>
<td>MTE</td>
<td>97</td>
<td>813</td>
</tr>
<tr>
<td>SynchroT.</td>
<td>407</td>
<td>1505</td>
</tr>
<tr>
<td>Similis</td>
<td>337</td>
<td>405</td>
</tr>
<tr>
<td>Characteristics and sub-characteristics to be measured</td>
<td>METRICS</td>
<td></td>
</tr>
<tr>
<td>-------------------------------------------------------</td>
<td>---------</td>
<td></td>
</tr>
<tr>
<td><strong>EFFECTIVENESS</strong> Value between 0 (minimum) and 5 (maximum)</td>
<td>(EFE1+EFE2+EFE3)/3</td>
<td></td>
</tr>
<tr>
<td>EFE1.- Degree of accuracy – precision of tasks &amp; results</td>
<td>(P1+P7+P13+P19+P25+P31)/6</td>
<td></td>
</tr>
<tr>
<td>EFE2.- Degree of completeness (tasks are accomplished and results are not missing)</td>
<td>(P2+P8+P14+P20+P26+P32)/6</td>
<td></td>
</tr>
<tr>
<td>EFE3.- Frequency of errors</td>
<td>(P3+P9+P15+P21+P27+P33)/6</td>
<td></td>
</tr>
<tr>
<td><strong>EFICIENCY</strong> Value between 0 (minimum) and 5 (maximum)</td>
<td>(EFI1+EFI2+EFI3+EFI4)/3</td>
<td></td>
</tr>
<tr>
<td>EFI1.- Time spent in the accomplishment of the task.</td>
<td>(TM1+TM2+TM3+TM4+TM5+TM6)</td>
<td></td>
</tr>
<tr>
<td>EFI2.- Need to use additional sources (material, software, etc.) for the task</td>
<td>(P4+P10+P16+P22+P28+P34)/6</td>
<td></td>
</tr>
<tr>
<td>EFI3.- Productivity – effort exerted by the user to carry out the task</td>
<td>(P5+P11+P17+P23+P29+P35)/6</td>
<td></td>
</tr>
<tr>
<td>EFI4.- Need to consult the software Help to perform the task</td>
<td>(P6+P12+P18+P24+P30+P36)/6</td>
<td></td>
</tr>
<tr>
<td><strong>SATISFACTION</strong> Value between 0 (minimum) and 5 (maximum)</td>
<td>(P37+P38+P39)/3</td>
<td></td>
</tr>
<tr>
<td>SAT1.- Usefulness</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAT2.- Trust</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAT3.- Pleasure</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>CONTEXT COVERAGE</strong> Value between 0 (minimum) and 5 (maximum)</td>
<td>(P40+P41+P42)/3</td>
<td></td>
</tr>
<tr>
<td>COB1.- Context of use</td>
<td></td>
<td></td>
</tr>
<tr>
<td>COB2.- Flexibility</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# QUESTIONNAIRE

42 questions grouped by tasks

<table>
<thead>
<tr>
<th>TASK 1 - CONFIGURATION</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>P1</td>
<td>I could configure appropriately.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T2</td>
<td>P2</td>
<td>I had fully completed the setup.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T3-</td>
<td>P3</td>
<td>The extractor reported errors.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T4</td>
<td>P4</td>
<td>I had to consult external material.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T5</td>
<td>P5</td>
<td>Configuration took me a lot of time.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T6</td>
<td>P6</td>
<td>I had to consult the extractor help.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TASK 2 - IMPORT</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>T7</td>
<td>P7</td>
<td>I imported the required files.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T8</td>
<td>P8</td>
<td>The import task was fully completed.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T9</td>
<td>P9</td>
<td>The import was successful and accurate.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T10</td>
<td>P10</td>
<td>It was not necessary to consult the extractor help.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T11</td>
<td>P11</td>
<td>The file import took less effort.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T12</td>
<td>P12</td>
<td>I completed the import within the expected time.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TASK 3 - EXTRACTION</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>T13</td>
<td>P13</td>
<td>The extractor carried out the task.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T14</td>
<td>P14</td>
<td>The extractor managed to work.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T15</td>
<td>P15</td>
<td>The extraction generated a list of terms.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T16</td>
<td>P16</td>
<td>I needed to consult additional documentation.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T17</td>
<td>P17</td>
<td>I finished the extraction, but...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T18</td>
<td>P18</td>
<td>I did not have to consult the extractor help.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TASK 4 - VALIDATION</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>T19</td>
<td>P19</td>
<td>I could perform the validation.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T20</td>
<td>P20</td>
<td>The validation task was fully completed.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T21</td>
<td>P21</td>
<td>During the validation process, I...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T22</td>
<td>P22</td>
<td>It was not necessary to consult the extractor help.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T23</td>
<td>P23</td>
<td>Validation was easy and it took no more effort than expected.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T24</td>
<td>P24</td>
<td>I had to consult the extractor help during the validation process.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The statements having the (X) mark have been formulated differently to the initial score to avoid the acquiescence of respondents, so they will need to be reversed for the final score.!
CONCLUSIONS

• Managing terminology still takes a lot of time and effort, even in this increasingly computerized profession.

• Research on automatic terminology extraction has been around for more than 20 years and significant enhancements concerning bilingual extraction and bilingual corpora exploitation have been introduced.

• I briefly described the BATE under evaluation and illustrated some results obtained for accuracy and with the QIU model.

• Results make it clear that much more work has to be done for BATE to be considered of real help to translators and terminologists, mainly due to poor accuracy results.
References


Many thanks for your attention

Chelo Vargas-Sierra

University of Alicante
IULMA
Campus de San Vicente
Apdo. 99
03080 Alicante

Phone & Fax
Direct Line: +34 965903438
Fax: +34 965903800
chelo.vargas@ua.es

Social Media
@chelovargas