



TMX and SRX

Exchanging TM Data

LRC-X Conference 13-14 September 2005

Angelika Zerfass
Consultant and Trainer for Translation Tools



Agenda

- What is TMX
 - TMX levels
 - TMX examples
- After the data exchange...
- What is SRX



What is TMX

- It is an XML representation of translation memory data
 - Header
 - Body

<header

creationtool="Déjà Vu "	—————	Déjà Vu, Transit, Trados, SDLX
creationtoolversion="4"	—————	Version / build number of the tool
datatype="PlainText"	—————	HTML, SGML, RTF, Interleaf, Java...
segtype="sentence"	—————	Basic segmentation
adminlang="en-us"	—————	Default language for elements like <note>
srclang="en-us"	—————	Source text language
o-tmf="DVMDB"	—————	Original translation memory format (DVMDB – Déjà Vu database...)

>



What is TMX

– Body

```
<body>
  <tu creationdate="20030915T153704Z" creationid="USER">
    <tuv lang="EN-US">
      <seg>This is the first sentence.</seg>
    </tuv>
    <tuv lang="DE-DE">
      <seg>Dies ist der erste Satz</seg>
    </tuv>
  </tu>
</body>
```

tu = Translation Unit, tuv lang = translation unit variant (language),
seg = segment



What is TMX

- Depending on the tool that created the TMX file, it can be bilingual or multilingual.
- Importing multilingual TMX file into a bilingual project will only import the relevant languages

```
<tu tuid="1" datatype="Text" srclang="en-us">
  <tuv xml:lang="en-us">
    <seg>This is a test.</seg>
  </tuv>
  <tuv xml:lang="de">
    <seg>Dies ist ein Test.</seg>
  </tuv>
  <tuv xml:lang="ja">
    <seg>テストです。</seg>
  </tuv>
  <tuv xml:lang="zh-cn">
    <seg>这是试验。</seg>
  </tuv>
</tu>
```



Levels of TMX

● Level 1:

- Plain text only (sufficient for data coming from software localization tools)

● Level 2:

- Text plus formatting (data coming from translation memory tools used for translation of documentation)

● To move formatting and text from one tool to the other both tools need to be level 2 compliant!



Level 1

- Formatting that is applied to the source and target text of a translation unit is not exported to the TMX file, only pure text.
- Original
 - This **sentence** has *some* formatting.
- In TMX
 - This sentence has some formatting.



Level 2

- Formatting that is applied to the source and target text of a translation unit is exported to the TMX file.
- Different tools use different ways of encoding that information.



TMX from Déjà Vu (Atril)

• Original

- This **sentence** contains *different* formatting information.

• In TMX from Déjà Vu

- `<seg>`
`<ph x="1">{1}</ph>` This
`<ph x="2">{2}</ph>` sentence
`<ph x="3">{3}</ph>` contains
`<ph x="4">{4}</ph>` different
`<ph x="5">{5}</ph><ph x="6">{6}</ph>` formatting information
`<ph x="7">{7}</ph>`.
`</seg>`
- DV puts placeholders (ph) where the formatting will go, not the formatting information itself, formatting information is stored in a separate file.



TMX from Translator's Workbench (Trados)

Original

- This **sentence** contains *different* formatting information.

In TMX from Translator's Workbench

- `<seg>`
This `<ut>{\b /ut>sentence<ut>}</ut>` contains
`<ut>{\i </ut>different<ut>}</ut>`
`<ut>{\ul </ut>formatting information<ut>}</ut>`.
`</seg>`
- `<seg>`
This
`<bpt i="1" type="bold">{\b </bpt>sentence<ept i="1">}</ept>`
contains
`<bpt i="2" type="italic">{\i </bpt>different<ept i="2">}</ept>`
`<bpt i="3">{\ul </bpt>formatting information<ept i="3">}</ept>`.
`</seg>`

- Example 1 is from Version 6.5, example 2 from version 7



TMX from Transit (Star)

● Original

- This **sentence** contains *different* formatting information.

● In TMX from Transit

- `<seg>`
This
`<bpt i="1" x="1" type="bold"></bpt> sentence<ept i="1"></ept>`
contains
`<bpt i="2" x="2" type="italic"><i></bpt>different<ept i="2"></i></ept>`
`<bpt i="3" x="3" type="ulined"><u options="single"></bpt>`
formatting information`<epti="3"></u></ept>`.
`</seg>`
- Transit uses the begin paired tag (bpt) the end paired tag (ept) and the information for bold (b), italics (i) and underlined (u)



TMX from SDLX (SDL)

● Original

- This **sentence** contains *different* formatting information.

● In TMX from SDLX

- `<seg>`
This
`<bpt i="1"x="1"><1></bpt>`sentence
`<epti="1"></1></ept>`
contains
`<bpt i="2"x="2"><2></bpt>`different `<epti="2">`
`</2></ept>`
`<bpt i="3"x="3"><3></bpt>`
formatting information`<epti="3"></3></ept>`.
`</seg>`
- SDLX uses placeholders for formatting information that is stored in a different file



Implications of different tags for formatting

- Tools that use placeholder tags do not include the actual formatting information in the TMX file
 - Other tools can only re-use the text
 - The result of the exchange is the same as with TMX level 1 (text only)
- TMX files which carry the actual formatting information will yield better matches in other tools that can read this information.



TMX specification

- TMX is a recommendation by OSCAR
 - OSCAR: LISA special interest group
 - Open Standards for Container/Content Allowing Re-use
 - The latest specification can be downloaded from <http://www.lisa.org/tmx/tmx.htm>
 - For comments: tmx@lisa.org
 - List of TMX certified tools
- The purpose of the TMX format is to provide a standard method to describe translation memory data that is being exchanged among tools and/or translation vendors, while introducing little or no loss of critical data during the process.



Does it work?

- With the current versions of translation tools on the market it works quite well
 - Previous versions sometimes created their own “flavor” of TMX which could not readily be imported by other tools, but the export files had to be changed before import. (en-us, EN_US)
- Yes, it does what it was developed for, it makes the exchange of data between tools possible...
- BUT - This is only half of the story...
- The question is, how well can the data that has been exchanged be used...



Reusing TMX data

- Although Translation Memory Tools have the same basic idea (storing source-target language pairs and recycling translations), this has been realized in different ways.
- Main issue here, are the segmentation rules



Segmentation rules

- Rules that the tool applies to the text to translate to split it up into segments
 - paragraph
 - sentence
 - phrase
 - incomplete sentences in bulleted lists
 - single words (headings, “Note”, “Attention”)



Segmentation rules

- End of segment rules (common to the default settings of all tools)
 - Dot at the end of a sentence (not after known abbreviations)
 - Question mark, exclamation mark
 - Paragraph mark
 - Colon
- End of segment rules (different for different tools)
 - Semicolon
 - Tab character
 - Sub segments (index entries, footnotes, graphics)



Comparison of default rules

	Workbench	Transit	DV	SDLX	Across
Colon	end	end	end	no end	no end
Semi-colon	no end	end	end	no end	no end
Tab	end	no end	no end	no end	no end
Soft return	no end	no end	end in Word no end in PPT	end in Word no end in PPT	no end



Example: semicolon

• Tool A

- Semicolon is end of segment
 - This is a sentence; this is another sentence.
- TM system sees two separate segments

• Tool B

- Semicolon is NOT end of segment
 - This is a sentence; this is another sentence.
- TM system sees one segment
- No match from the TMX data!
 - Match rate around 50%, usual setting around 70%



Settings for better reuse...

- Check the segmentation settings of the source tool, if possible
- Re-create this setting in the target tool, as far as possible
- Set down the minimum match value from the default 75% to about 50%
- For TM data that does not yield useful results, you may have to run an alignment of the original material on the target system.



Next step - SRX

- Segmentation Rules Exchange
- When exporting TM data to a TMX file, the segmentation rules are written to an extra file.
- If the receiving system is able to create the same setting as the TMX-exporting system, the recycling rates for matches will get better.



SRX

- SRX is under developed at the moment. The SRX file will contain the following information:
 - `<languagerules>` - Definition of the rules of a specific language
 - `<maprules>` - Definition, how those rules were set at the time of the TMX export



Endrules and exceptions

● Rule:

- A dot followed by a space is the end of a segment..

- This is the first sentence. This is the second sentence.

● Exception:

- A dot, preceded by a number is not the end of a segment.

- Dies ist der 1. Satz.



End rules and exceptions

● Rule:

- ...
<endrule>
<beforebreak> [\.]</beforebreak>
<afterbreak>\s</afterbreak>
<excludeexception exceptionname="numbers">
</endrule>
...

● Exception for numbers, abbreviations...

- ...
<exception exceptionname="numbers">
<beforebreak>[0-9]+\.<beforebreak>
<afterbreak>\s</afterbreak>
</exception>
...



What can SRX not do?

- It can only show the segmentation rule settings at the time of export.
- It cannot show any changes that have been applied in the segmentation rules during the use of the TM.
- Sometimes the rules from system 1 cannot be re-created in system 2, then the rule will be ignored.



SRX Specification

• Latest version

– www.lisa.org/srx/srx.htm

– www.lisa.org/srx/srx10-20040420.htm



Next level

- Up to now the tools can only exchange information on text and its formatting.
- SRX will come soon.
- Next level after that would be the exchange of additional data like project name, customer name...but as the tools differ very much in what they offer this will be difficult
 - Some tools offer free creation of fieldnames others only offer a certain set of fields



TMX discussion lists

- http://groups.yahoo.com/group/tmx_software/
 - For TMX developers, founded July 2003, less than 5 members, seems to have very low traffic
- <http://groups.yahoo.com/group/DataDefinition/>
 - founded November 2000, 190 members
 - Localization Clients and Vendors looking at standards together so that we can standardize on a Translation Object.
 - Examining OPENTAG, TMX and other XML standards.
- http://groups.yahoo.com/group/tmx_lisa/
 - Translation Memory Exchange Standards Mailing List
Mailing list to discuss TMX and other related standards. Said to have very low traffic.
- <http://www.lisa.org/tmx/>
 - TMX implementation mailing list



Thank you
for your attention

Any Questions?

Angelika Zerfass