



TMX

Austauschformat für Translation Memory Systeme

2. Kölner Tagung zum Thema
Softwarelokalisierung
11.-12. März 2005

Agenda

- Einführung in TMX
- Was kann TMX, was kann es nicht
- Nächster Schritt: SRX

Was ist TMX

- Translation Memory Exchange Format
- Eine XML Darstellung der Inhalte eines TMs (Translation Memory)
 - Header
 - Body

<header

creationtool="Déjà Vu "	—————	Déjà Vu, Transit, Trados, SDLX
creationtoolversion="4"	—————	Version / Build Nummer des Tools
datatype="PlainText"	—————	HTML, SGML, RTF, Interleaf, Java...
segtype="sentence"	—————	Segmentierung
adminlang="en-us"	—————	Standardsprache für Elemente wie <note>
srclang="en-us"	—————	Sprache des Ausgangstextes
o-tmf="DVMDB"	—————	Originalformat des TM Systems (DVMDB – Déjà Vu database...)

>

Was ist 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

Was ist TMX

- Je nach Tool, welches die TMX Datei erstellt, kann diese bilingual oder multilingual sein.
- Import einer multilingualen TMX Datei in ein bilinguales System importiert nur die relevanten Sprachen.

```
<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>
```

TMX Stufen

- Stufe 1:
 - Nur Text (z.B. ausreichend für Daten eines Software-Lokalisierungstools)
- Stufe 2:
 - Text plus Formatierung (wird von den meisten Translation Memory Systemen verwendet)
- Um Text und Formatierung auszutauschen zu können, müssen beide Systeme – das TMX-exportierende und das TMX-importierende den Austausch auf Stufe 2 beherrschen.

Stufe 1

- Formatierungsinformationen werde nicht in die TMX Datei exportiert.
- Original
 - Dieser **Satz** enthält *etwas* Formatierung.
- In TMX
 - Dieser Satz enthält etwas Formatierung.

Stufe 2

- Formatierungsinformation der Segmente in Ausgangs- und Zielsprache wird in die TMX Datei exportiert.
- Verschiedene Systeme verwenden unterschiedliche Arten der Kodierung.

TMX aus Déjà Vu (Atril)

- Original
 - This **sentence** contains *different* formatting information.
- In TMX aus 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 setzt Platzhalter (ph) für die Formatierungsinformation ein, nicht die Information, welche Formatierung hier verwendet wurde.
 - Beim Import dieser Datei in ein anderes System geht die Formatierungsinformation verloren.

TMX aus Translator's Workbench (Trados)

- Original
 - This **sentence** contains *different formatting information*.
- In TMX aus 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>`
 - Trados verwendet das Tag (ut = unknown tag) sowie die Information (b) für Fettdruck, (i) für Kursivdruck und (ul) für Unterstreichung
 - Diese Information wird, je nach empfangendem System, hohe Fuzzy-Matches produzieren.

TMX aus 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 verwendet das Tagpaar (bpt=begin paired tag) und (ept=end paired tag) sowie die Information (b) für Fettdruck, (i) für Kursivdruck und (ul) für Unterstreichung
 - Diese Information wird, je nach empfangendem System, hohe Fuzzy-Matches produzieren.

TMX aus SDLX (SDL)

- Original
 - This **sentence** contains *different* formatting information.
- In TMX aus 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 verwendet Platzhalter anstelle der eigentlichen Formatierungsinformation
 - Beim Import dieser Datei in ein anderes System geht die Formatierungsinformation verloren.

Wie gut funktioniert es?

- Ja, TMX erfüllt die Spezifikation: Mit den aktuell verfügbaren TM Systemen funktioniert der TMX Austausch
 - 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.
- Aber....die Frage ist, wie gut lassen sich die ausgetauschten Inhalte wiederverwenden.

Wiederverwendbarkeit

- Obwohl alle TM Systeme die gleiche Grundidee haben, nämlich Segmentpaare abzuspeichern, sieht jedes System die Segmente eines Textes anders.
- Hauptunterschied:
Segmentierungsregeln

Segmentierungsregeln

- Regeln, nach denen ein Text in Segmente aufgespalten wird.
 - Abschnitt
 - Satz
 - Teilsatz in Aufzählungslisten
 - Untersegmente wie Fußnoten oder Indexeinträge
 - Einzelstehende Wörter (Achtung:)

Segmentierungsregeln

- Endregeln (in den Regellisten aller Systeme gleich)
 - Satzendezeichen (nicht bei bekannten Abkürzungen)
 - Fragezeichen, Ausrufezeichen
 - Absatzmarke
 - Doppelpunkt
- Endregeln (unterschiedlich je nach System)
 - Strichpunkt
 - Tabulator
 - Untersegmente (Indexeinträge, Fußnoten, Grafiken)

Vergleich der Standardregeln

	Workbench	Transit	DV	SDLX	Across
Doppelpunkt	Ende	Ende	Ende	Kein Ende	Kein Ende
Strichpunkt	Kein Ende	Ende	Ende	Kein Ende	Kein Ende
Tabulator	Ende	Kein Ende	Kein Ende	Kein Ende	Kein Ende
Bedingte Zeilenschaltung/ Soft return	Kein Ende	Kein Ende	Ende in Word, Kein Ende in PPT	Ende in Word, Kein Ende in PPT	Kein Ende

Beispiel Strichpunkt

- System A
 - Strichpunkt ist Ende eines Segments
 - Dies ist ein Satz; dies ist ein weiterer Satz.
 - TM System sieht zwei separate Segmente
- System B
 - Strichpunkt ist NICHT Ende eines Segments
 - Dies ist ein Satz; dies ist ein weiterer Satz.
 - TM System sieht ein Segment
 - Kein Match aus den TMX Daten!
 - Einstellung der minimalen Matchrate der TM Systeme steht häufig bei ca. 70%. Hier würde eine Einstellung von 50% gebraucht.

Nächste Stufe - SRX

- Segmentation Rules Exchange
 - Beim Austausch der TM Daten über TMX werden nicht nur die Segmente mit/ohne Formatierung transferiert, sondern auch Informationen zu den Segmentierungsregeln und deren Ausnahmen.

Was kann SRX nicht?

- Es werden nur die Informationen zu den aktuellen Segmentierungsregeln des exportierenden Systems weitergegeben.
- Informationen zu geänderten Regeln werden nicht weitergegeben.
- Manchmal können die Regeln von System 1 nicht in System 2 nachgebildet werden, hier geht die Information verloren.

TMX Spezifikation

- TMX wird durch die OSCAR Gruppe der LISA bearbeitet
 - OSCAR: Open Standards for Container/Content Allowing Re-use
 - Aktuelle Version
<http://www.lisa.org/tmx/tmx.htm>
 - Kommentare an: tmx@lisa.org

SRX Spezifikation

- Letzte Version
 - www.lisa.org/srx/srx.htm
 - www.lisa.org/srx/srx03-20030724.htm



z a a c

Angelika Zerfaß

Thank you