# Standards basés sur XML pour la traduction et la localisation

Andrei Popescu-Belis
TIM / ETI, Université de Genève

*Cours n°12*

## Plan du cours

1. Format d'échange pour les mémoires de traduction : TMX
2. Encodage normalisé des conventions de segmentation : SRX
3. Format XML des données pour la localisation des programmes : XLIFF



## TMX : Translation Memory eXchange

- Définition
  - format de stockage et d'échange des mémoires de traduction
  - indépendant d'un système particulier
  - « open-source », basé sur XML
- Application
  - mémoires de traduction (e.g., Trados, SDLX)
  - outils de localisation
- Objectifs
  - faciliter l'échange des données entre les outils ou entre les professionnels de la traduction
  - réduire le monopole des grands systèmes en permettant le transfert des mémoires

http://www.lisa.org/standards/tmx




## Organisation

- TMX conçu par le groupe OSCAR
  - **O**pen **S**tandards for **C**ontainer/Content **A**llowing **R**e-use
  - un Special Interest Group de LISA
  - LISA = Localisation Industry Standards Assocation
- Membres du comité de pilotage d'OSCAR
  - représentants de :
    - Brigham Young University, GlobalSight, IBM, JDEdwards, RWS Group, Sun Microsystems, Star Group, SDL, SAP, Trados, Welocalize
  - assurance que les « grands » systèmes utilisent TMX



## Etat actuel

- Version actuelle du standard: 1.4 b, avril 2004
- Diffusion du standard
  - plusieurs systèmes implémentent TMX, dont TRADOS, Transit, SDLX, Déjà Vu
    - mais tout le standard n'est pas toujours respecté
  - plusieurs systèmes sont « certifiés TMX »
    - TRADOS 7 (depuis 2005), SDLX 2004 et 2003 (SDL International, depuis 2004), WorldServer (Idiom Technologies, depuis 2005), Ambassador (GlobalSight, depuis 2004)
  - importation/exportation de mémoires de traduction
    - chaque logiciel utilise pour le stockage son format propriétaire



## Aperçu du format sur un exemple (1)

```
<?xml version="1.0"?>
<!-- Example of TMX document -->
<tmx version="1.4">
 <header      creationtool="XYZTool"
              creationtoolversion="1.01-023"
              datatype="PlainText"
              segtype="sentence"
              adminlang="en-us"
              srclang="EN"
              o-tmf="ABCTransMem"
              creationdate="20020101T163812Z"
              creationid="ThomasJ"
              changedate="20020413T023401Z"
              changeid="Amity"
              o-encoding="iso-8859-1">
 <note>This is a note at document level.</note>
 <prop type="RTFPreamble">{\rtf1\ansi\tag etc...{\fonttbl}</prop>

 <ude name="MacRoman" base="Macintosh">
     <map unicode="#xF8FF" code="#xF0" ent="Apple_logo" subst="[Apple]"/>
 </ude>
 </header>
```

## Aperçu du format sur un exemple (2)

```
<body>
 <tu tuid="0001" datatype="Text"
    usagecount="2"
    lastusagedate="19970314T023401Z">

  <note>Text of a note at the TU
    level.</note>
  <prop type="x-
    Domain">Computing</prop>
  <prop type="x-
    Project">P&#x00E6;gasus</prop>
  <tuv  xml:lang="EN"
    creationdate="19970212T153400Z"
    creationid="BobW">

   <seg>data (with a non-standard
    character: &#xF8FF;).</seg>

  </tuv>
```

```
  <tuv xml:lang="FR-CA"
    creationdate="19970309T021145Z"
    creationid="BobW"
    changedate="19970314T023401Z"
    changeid="ManonD">

    <prop type="Origin">MT</prop>

    <seg>donn&#xE9;es (avec un
      caract&#xE8;re non standard:
      &#xF8FF;).</seg>

  </tuv>
 </tu></body></tmx>
```



## Aperçu du format sur un exemple (3)

```
    <tu  tuid="0002" srclang="*all*">

        <prop type="Domain">Cooking</prop>
        <tuv xml:lang="EN">
            <seg>menu</seg>
        </tuv>
        <tuv xml:lang="FR-CA">
            <seg>menu</seg>
        </tuv>
        <tuv xml:lang="FR-FR">
            <seg>menu</seg>
        </tuv>

    </tu>
</body>
</tmx>
```



## Exemple 2 : localisation d'un manuel

- Supposons qu'un manuel technique contienne la phrase:
  - « If you click the  button, APPLICATION will not respond; instead it displays the following message: *Application error. Please ask for further instruction.* »
- Noter la présence :
  - d'un bouton (objet non linguistique)
  - du formatage interne à un segment : le mot APPLICATION
  - du formatage recouvrant deux segments : le message d'erreur en italiques

Source : *Multilingual Computing & Technology*, **14**(2), 2003.



## Stockage de l'exemple dans une mémoire de traduction

- **Objectif** : stocker le maximum d'information pour avoir des 'match' à 100% sans perdre ou déformer le formatage graphique et les codes
- Utilisation des balises TMX par les systèmes : explications
  - <tu> : translation unit, contient des <tuv> (t.u. variants)
  - <ph> : place holder, contient du code non linguistique indépendant
  - <bpt> : begin paired tag, code non linguistique, balise ouvrante
  - <ept> : end paired tag, code non linguistique, balise fermante
  - <it> : isolated tag, code non linguistique, balise ouvrante/fermante isolée dans un segment (l'autre balise étant dans un segment différent)
  - <ut> : unknown tag
- NOTE : puisque TMX est du XML, '**<**' et '**>**' en dehors des balises, sont remplacés par **<** et **>** – sauf dans ce qui suit, pour la lisibilité



## TRADOS 5.5 – TMX v. 1.1

```
<tuv lang="EN-US">
    <seg>
        If you click the
        <ut>{\pict}</ut>
        button,
        <ut>{\scaps </ut>
        Application
        <ut>}</ut>
        will not respond; instead it displays the following message:
        <ut>{\i </ut>
        Application error.
        <ut>}</ut>
    </seg>
</tuv>
```



## Transit XV SP1 – TMX v. 1.1

```
<tuv lang="en-us">
    <seg>
        <ph type="image"><object id="0" type="unknown" amount="9"/></ph>
        If you click the
        <ph type="image"><object id="1" type="picture" amount="2"/></ph>
        button,
        <ph type="image"><object id="13" type="unknown"/></ph>
        <bpt i="1" type="font"><F id="1"></bpt>
        Application
        <ph type="image"><object id="14" type="unknown"/></ph>
        <ept i="1"></F></ept>
        will not respond; instead it displays the following message:
        <ph type="image"><object id="15" type="unknown"/></ph>
        <bpt i="2" type="italic"><i></bpt>
        Application error.
        <ept i="2"></i></ept>
    </seg>
</tuv>
```

## SDLX 4.2.1 – TMX v. 1.4

```
<tuv xml:lang="EN-US">
    <seg>
        <bpt i="1" x="1"><1></bpt>
        If you click the
        <ept i="1"></1></ept>
        <ph x="2"><2/></ph>
        <bpt i="2" x="3"><3></bpt>
        button,
        <ept i="2"></3></ept>
        <bpt i="3" x="4"><4></bpt>
        Application
        <ept i="3"></4></ept>
        <bpt i="4" x="5"><5></bpt>
        will not respond; instead it displays the following message:
        <ept i="4"></5></ept>
        <bpt i="5" x="6"><6></bpt>
        Application error
        <ept i="5"></6></ept>
        <it pos="begin" x="7"><7></it> .
    </seg>
</tuv>
```



## Déjà Vu X – TMX v. 1.4

```
<tuv xml:lang="en-us">
    <seg>
        If you click the
        <ph x="1">{1}</ph>
        button,
        <ph x="2">{2}</ph>
        Application
        <ph x="3">{3}</ph>
        will not respond; instead it displays the following message:
        <ph x="4">{4}</ph>
        Application error
        <ph x="5">{5}</ph>.
        </seg>
</tuv>
```



## Vérification de la compatibilité d'un outil avec la norme TMX (1)

- Objectif : exporter des mémoires conformes à TMX
- Difficulté : le marquage éventuel entre les mots

- **Level 1** (texte simple)
  - seulement le contenu, pas l'information de formatage
  - ce niveau est en fait suffisant pour des textes qui n'ont pas de codes insérés (i.e. balises, instructions logicielles, etc.)
- **Level 2** (marquage du contenu)
  - contenu textuel + instructions de formatage
  - la conformité permet ici à d'autres outils de niveau 2 utilisant la mémoire de recréer le format original du document



## Vérification de la compatibilité d'un outil avec TMX (2)

- Note
  - la conformité ne tient pas compte du problème (difficile) de la segmentation
    - si deux logiciels segmentent différemment une phrase, même s'ils respectent TMX, le résultat pourra ne pas être le même en sortie avec la même mémoire de traduction

- Vérification de la conformité
  - outils disponibles sur Internet
    - http://www.lisa.org/standards/tmx/certification.html
  - mais la certification officielle est délivrée seulement par un laboratoire habilité par la LISA



## Comment vérifier soi-même la conformité d'un logiciel à TMX ?

- Le format d'exportation et d'importation d'un logiciel de mémoire de traduction doit respecter le standard TMX, basé sur XML

- Principe
  - choisir un logiciel de TM
  - effectuer une suite de tests
    - construire des mémoires de traduction simples
    - importation, exécution, comparaison du résultat obtenu avec le résultat attendu
    - idem pour l'exportation



## Procédure de vérification

- Télécharger le TMX Compliance kit à partir de :
  - http://www.lisa.org/tmx/specification.html
- Le désarchiver et examiner les fichiers et la documentation
  - spécification TMX (notamment la DTD)
  - documentation de la procédure de vérification de la conformité à TMX
- Appliquer point par point la procédure et noter à chaque étape le résultat du test de conformité
  - utiliser le programme fourni pour la vérification/validation/comparaison

## Plan du cours

1. Format d'échange pour les mémoires de traduction : TMX
2. Encodage normalisé des conventions de segmentation : SRX
3. Format XML des données pour la localisation des programmes : XLIFF

19

## SRX : Segmentation Rules eXchange

- Version 1.0
  - avril 2004 : officiellement acceptée comme norme de LISA/OSCAR
  - http://www.lisa.org/standards/srx
- Peut-être en train d'être remplacée par une norme de l'ISO ?
  - http://www.tc37sc4.org/WG2/wg2.htm
  - documents en cours d'élaboration…

20

## Aperçu de SRX (1)

- SRX complémente la norme TMX
  - permet de rendre explicites les règles de segmentation de phrases utilisées pour créer une mémoire de traduction
  - faciliter l'échange de mémoires de traduction
- Deux parties
  - les règles de segmentation
  - les correspondances langues/règles (*simple*)
- Règles de segmentation
  - utilisent les « expressions régulières »
  - intuitivement : expressions avec « jokers »

21

## Aperçu de SRX (2)

- Document TMX + document SRX
  - permettent d'expliquer comment le texte a été segmenté avant d'être introduit dans une mémoire de traduction
- État actuel : segmentation en phrases
  - les mémoires de traduction courantes sont aussi basées sur les phrases
    - surtout celles qui gèrent le format TMX
  - à l'avenir : segmentation plus complexe, en syntagmes et termes
- SRX est implémenté en XML
  - les documents SRX sont bien formés et valides
  - la norme fournit une DTD et un schéma

22

**Texte à segmenter** :
The U.K. Prime Minister, Mr. Blair, was seen out with his family today.

| Rule set | Result | Notes |
|---|---|---|
| `<rule break="yes">`<br>`  <beforebreak>[\.\?!]+</beforebreak>`<br>`  <afterbreak>\s</afterbreak>`<br>`</rule>` | (1) The U.K.<br>(2) Prime Minister, Mr.<br>(3) Blair, was seen out with his family today | The simple full-stop followed by a space rule here showing its limitations |
| `<rule break="no">`<br>`  <beforebreak>U.K.</beforebreak>`<br>`  <afterbreak>\s</afterbreak>`<br>`</rule>`<br>`<rule break="yes">`<br>`  <beforebreak>[\.\?!]+</beforebreak>`<br>`  <afterbreak>\s</afterbreak>`<br>`</rule>` | (1) The U.K. Prime Minister, Mr.<br>(2) Blair, was seen out with his family today | Partially corrected with an exception for "U.K." |
| `<rule break="no">`<br>`  <beforebreak>U.K.</beforebreak>`<br>`  <afterbreak>\s</afterbreak>`<br>`</rule>`<br>`<rule break="no">`<br>`  <beforebreak>Mr.</beforebreak>`<br>`  <afterbreak>\s</afterbreak>`<br>`</rule>`<br>`<rule break="yes">`<br>`  <beforebreak>[\.\?!]+</beforebreak>`<br>`  <afterbreak>\s</afterbreak>`<br>`</rule>` | (1) The U.K. Prime Minister, Mr. Blair, was seen out with his family today | Sufficient exceptions to prevent segmentation on "U.K." and "Mr." |

23

## Plan du cours

1. Format d'échange pour les mémoires de traduction : TMX
2. Encodage normalisé des conventions de segmentation : SRX
3. Format XML des données pour la localisation des programmes : XLIFF

24

## Utilisation de TMX pour la localisation

- Séparation langue *vs.* code en TMX 1.4b
- Exemple : élément d'un document HTML avec du texte dans une valeur d'attribut

```
See the <A TITLE="Go to Notes" HREF="notes.htm">Notes</A>
for more details.
```

- Codage en TMX avec marquage du contenu

```
See the  <bpt i="1" type="link">&lt;A TITLE="<sub> Go to
Notes</sub> " HREF="notes.htm"></bpt> Notes
<ept i="1">&lt;/A></ept> for more details.
```

- Séparation correcte, mais difficulté de préciser quels sont les segments à cause de l'insertion de "Go to Notes" – qui doit être traduit aussi

25

## OpenTag et XLIFF

- Objectifs
  - séparer, dans une application :
    - le texte qui doit être localisé (menus, boutons, aide, infos, etc.)
    - les éléments de programme qui ne doivent pas changer
  - fusionner ensuite la traduction avec le programme
- OpenTag et XLIFF
  - même objectif, mais XLIFF est plus précis & interopérable
  - "XML Localization Interchange File Format"
- Traduction du texte séparé : avec un outil qui ne change pas les balises (ex.: Trados TagEditor)

26

## OpenTag

- ***N'est plus mis à jour***, mais simple à comprendre
- Application capable d'extraire et de fusionner des fichiers
  - OpenTag définit seulement le format, pas la technique d'extraction
  - la technique dépend du format du fichier à localiser
- 1. Extraction
  - à partir du fichier initial
  - partie localisable : fichier OpenTag (.OTF)
  - partie non linguistique : fichier « squelette » (.SKL)
  - le SKL contient des liens vers les items correspondants dans le OTF
- 2. Traduction / localisation du OTF
- 3. Fusion
  - on obtient un programme localisé autonome

27

## XLIFF

- Objectifs
  - les mêmes que OpenTag, mais avec une inspiration de TMX
- Organisation
  - depuis 2001 : comité technique de OASIS
  - membres : Oracle, Novell, IBM/Lotus, Sun MicroSystems, Alchemy Software, Berlitz, Moravia-IT, et RWS Group
- Spécification
  - http://www.oasis-open.org/committees/xliff/documents/xliff-specification.htm
  - Version 1.0, « committee specification », début 2002
  - Version 1.1, « committee specification », octobre 2003
- Voir aussi
  - http://www.xliff.org
  - http://www.opentag.com/xliff.htm

28

## Différences avec OpenTag et TMX

- XLIFF *vs.* OpenTag : même objectif
  - OpenTag est plus ancien, XML était nouveau à l'époque
  - XLIFF est plus structuré
  - XLIFF peut incorporer le fichier « squelette » dans le fichier à localiser (autonomie)
  - XLIFF offre un mécanisme développé pour la pré-traduction, la révision, l'enregistrement des versions
- XLIFF
  - seulement deux langues dans chaque document
  - OpenTag et TMX : autant de langues qu'on veut
  - choix de fixer **deux** langues : pour simplifier le modèle
- XLIFF *vs.* TMX et OpenTag
  - TMX peut encapsuler du code-machine dans le fichier à traduire
  - OpenTag utilise un système d'ancres et de liens
  - XLIFF peut faire les deux

29

## Exemple A en XLIFF (1/2) [fichier squelette externe]

```
<?xml version="1.0"?>
<xliff version="1.0">
<file original="Graphic Example.psd" source-language="EN-US" target-
    language="JA-JP" tool="Rainbow" datatype="photoshop">
<header>
<skl>
    <external-file uid="3BB236513BB24732" href="Graphic Example.psd.skl"/>
</skl>
<phase-group>
    <phase phase-name="extract" process-name="extraction"
           tool="Rainbow" date="20010926T152258Z"
           company-name="NeverLand Inc." job-id="123"
           contact-name="Peter Pan" contact-email="ppan@xyzcorp.com">
     <note>Make sure to use the glossary I sent you yesterday. Thanks.</note>
    </phase>
 </phase-group>
</header>
…
```

30

## Exemple A en XLIFF (2/2)

```
…
<body>
<trans-unit id="1" maxbytes="14">
    <source xml:lang="EN-US">Quetzal</source>
    <target xml:lang="JA-JP">Quetzal</target>
</trans-unit>
<trans-unit id="3" maxbytes="114">
    <source xml:lang="EN-US">An application to manipulate and
                              process XLIFF documents</source>
    <target xml:lang="JA-JP">XLIFF 文書を編集、または処理するアプリケ
    ーションです。</target>
</trans-unit>
<trans-unit id="4" maxbytes="36">
    <source xml:lang="EN-US">XLIFF Data Manager</source>
    <target xml:lang="JA-JP">XLIFF データ・マネージャ</target>
</trans-unit>
</body> </file> </xliff>
```



## Exemple B en XLIFF (1/2) [fichier squelette interne]

```
<?xml version="1.0" encoding="windows-1252" ?>
<xliff version="1.0" xml:lang='en'>
    <file source-language='en' target-language='fr' datatype="winres" original="Sample1.rc">
    <header>
    <skl>
    <internal-file crc="64a2b9b0"><![CDATA[
<OKFSKL100:RES:964008261>
#include "resource.h"
IDD_DIALOG1 DIALOG DISCARDABLE  0, 0, 186, 57
STYLE DS_MODALFRAME | WS_POPUP | WS_CAPTION | WS_SYSMENU
CAPTION "<xref$1>"
FONT 8, "MS Sans Serif"
BEGIN
    LTEXT           "<xref$2>",IDC_STATIC,8,4,18,8
    EDITTEXT        IDC_EDIT1,8,16,100,14,ES_AUTOHSCROLL
    CONTROL         "<xref$3>",IDC_CHECK1,"Button",
                    BS_AUTOCHECKBOX | WS_GROUP |
                    WS_TABSTOP,8,40,41,10
    DEFPUSHBUTTON  "<xref$4>",IDOK,129,7,50,14,WS_GROUP
    PUSHBUTTON      "<xref$5>",IDCANCEL,129,24,50,14
END]]></internal-file>
    </skl>
    </header>
```



## Exemple B en XLIFF (2/2)

```
…
<body>
    <group restype="dialog" resname="IDD_DIALOG1">
            <trans-unit id="1" restype="caption">
                    <source>Title</source>
            </trans-unit>
            <trans-unit id="2" restype="label" resname="IDC_STATIC">
                    <source>&amp;Path:</source>
            </trans-unit>
            <trans-unit id="3" restype="check" resname="IDC_CHECK1">
                    <source>&amp;Validate</source>
            </trans-unit>
            <trans-unit id="4" restype="button" resname="IDOK">
                    <source>OK</source>
            </trans-unit>
            <trans-unit id="5" restype="button" resname="IDCANCEL">
                    <source>Cancel</source>
            </trans-unit>
    </group>
</body></file></xliff>
```



## Outils pour XLIFF

- Suivre les liens à www.opentag.com
  - XLIFF Settings Files (07/2001)
    - pour traduire des documents XLIFF avec SDLX ou TagEditor (http://www.opentag.com/downloads.htm)
  - outils Rainbow (famille d'outils RWS pour la localisation)
    - filtres XLIFF (extracteurs) pour différents formats de fichiers ressources
- La traduction de fichiers XLIFF simplifie le travail de localisation