MLIF: A Metamodel to Represent and Exchange Multilingual Textual Information

Samuel Cruz-Lara[1], Gil Francopoulo[2], Laurent Romary[3], Nasredine Semmar[4]
[1]INRIA-TALARIS, [2]Tagmatica, [3]INRIA-GEMO & HUB IDSL, [4]CEA-LIST
France
samuel.cruz-lara@loria.fr, gil.francopoulo@wanadoo.fr, laurent.romary@inria.fr, nasredine.semmar@cea.fr

Abstract

The fast evolution of language technology has produced pressing needs in standardization. The multiplicity of language resources representation levels and the specialization of these representations make difficult the interaction between linguistic resources and components manipulating these resources. In this paper, we describe the MultiLingual Information Framework (MLIF – ISO CD 24616). MLIF is a metamodel which allows the representation and the exchange of multilingual textual information. This generic metamodel is designed to provide a common platform for all the tools developed around the existing multilingual data exchange formats. This platform provides, on the one hand, a set of generic data categories for various application domains, and on the other hand, strategies for the interoperability with existing standards. The objective is to reach a better convergence between heterogeneous standardisation activities that are taking place in the domain of data modeling (XML; W3C), text management (TEI; TEIC), multilingual information (TMX-LISA; XLIFF-OASIS) and multimedia (SMILText; W3C). This is a work in progress within ISO-TC37 in order to define a new ISO standard.

Introduction

The scope of activities in localization and translation memory, as well as any type of online multilingual customization (subtitling, iTV, karaoke) is very large, and numerous independent groups are working on these aspects, such as LISA, OASIS, W3C, ISO, etc. Under the guidance of the above-mentioned groups, many formats have been developed. Some of the major formats of specific interest for localization and translation memories are TMX[1], XLIFF[2], ITS[3]. There are many identical requirements for all the formats irrespective of the differences in final output. For example, all the formats aim at being user-friendly, easy-to-learn, and at reusing existing databases or knowledge. All these formats work well in the specific field they are designed for, but they lack a synergy that would make them interoperable when using one type of information in a slightly different context. In this paper, we describe an empirical standardization work that will serve for proposing a metamodel and related data-categories to represent and exchange multilingual textual information.

1. MLIF metamodel

As with TMF[4] in terminology, MLIF (Cruz-Lara et al., 2008) provides a metamodel and a set of generic data categories for various application domains (Figure 1). MLIF describes not only the basic linguistic elements (sentence, syntactic component, word, part-of-speech, …), but can be used to represent the structure of the document (title, paragraph, section,…).

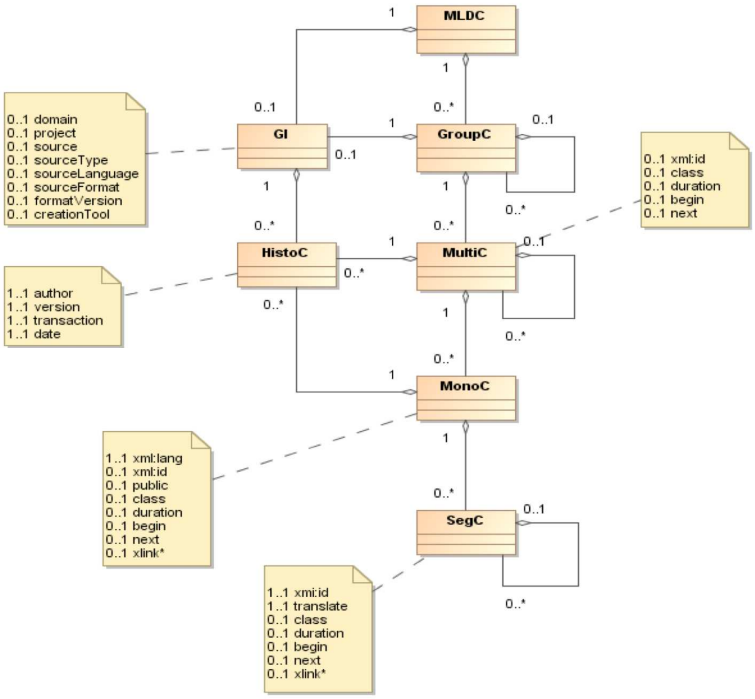

Figure 1: MLIF metamodel and its data categories

2. MLIF Applications

Because of the genericity of its metamodel and the facility to adapt data categories, MLIF can be used in several types of applications. The latest version of MLIF provides strategies for the interoperability and/or linking of models including, but not limited to: TMX, XLIFF, SMILText[5] and ITS. We present in the following sections some use cases of MLIF.

3.1 MLIF process with TMX

The purpose of TMX is to allow easier exchange of translation memory data between tools and/or translation vendors with little or no loss of critical data during the

[1] http://www.lisa.org/tmx
[2] http://www.oasis-open.org/committees/xliff/
[3] http://www.w3.org/TR/its/
[4] Computer application in terminology – Terminological Markup Framework, Geneva, International Organization for Standardization.
[5] http://www.w3.org/TR/SMIL3/smil-text.html.

process.

Figure 2 illustrates the interaction between TMX and MLIF. This process includes, in this order: extraction, translation, and merging. The starting point is a TMX document with linguistic content in English (en) and in German (de). The extraction process (1) gives first a "Skeleton File" (2) which contains all TM formatting information and secondly a MLIF file (3) in which only relevant linguistic information is stored. As most translators (human beings or automatic software modules) work with TMX software oriented-tools, a XSL style-sheet allows transforming a MLIF document into a TMX document or into any ML2 document. This file does not contain any formatting information. Once the translator adds the related Japanese translation, another XSL style-sheet allows transforming a TMX document into a MLIF document (4). Finally, the new MLIF document (containing the Japanese translation) is merged with the "Skeleton File" in order to obtain a new TMX formatted document (5).

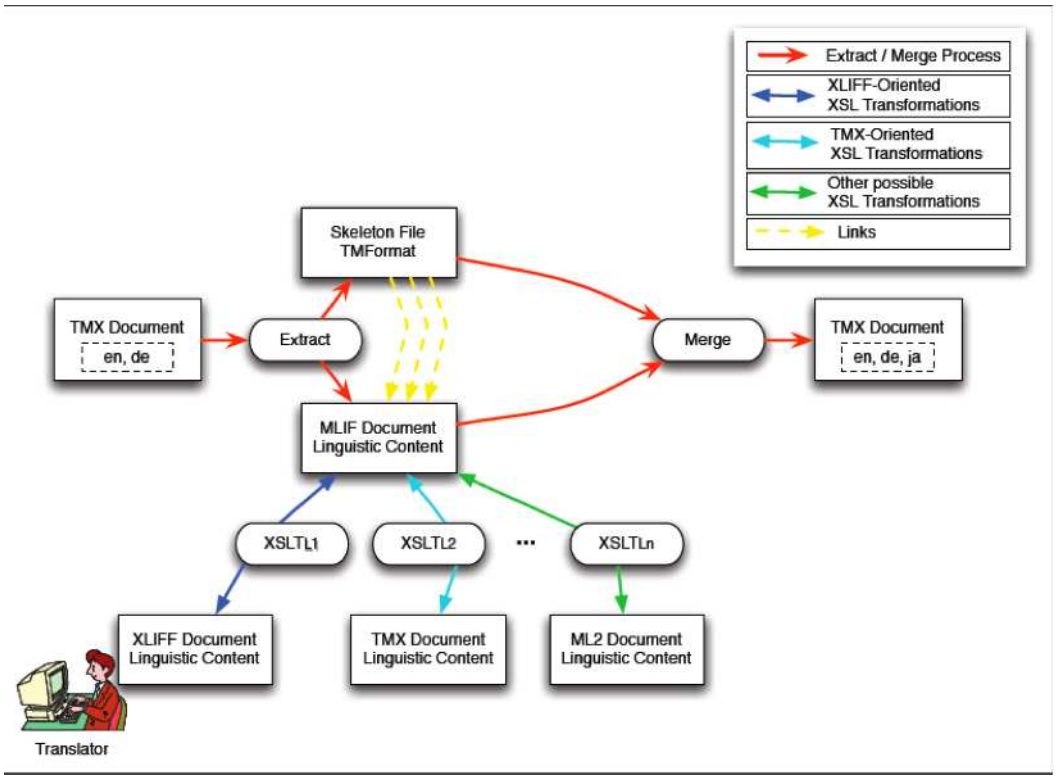

Figure 2: MLIF-TMX interaction.

3.2 XML details of a simple TMX application

The following example, based on TMX version 1.4, focuses on the multilingual units of a TMX document. It should be noted that TMX version 2.0 has been available as a Working Draft since March 28th, 2007.

```
<tmx version="1.4">
 <header
   adminlang="eng"
   creationdate="20040731T164933Z"
   creationtool="Heartsome TM Server"
   creationtoolversion="1.0.1"
   datatype="xml"
   o-tmf="unknown"
   segtype="block"
   srclang="*all*"/>

 <body>
  <tu creationdate="20020930T004233Z" tuid="1091303313515">
   <tuv xml:lang="fra">
    <seg>Le processus de contrôle de qualité en dix
```

```
          étapes qu'il a créé il y a plus de 1300 ans est
          beaucoup plus complet et précis que ceux existant
          aujourd'hui.</seg>
    </tuv>
    <tuv xml:lang="eng">
     <seg>His 10-stage quality control process initiated
          more than 1300 years ago is far more thorough and
          exacting than any existing today.</seg>
    </tuv>
   </tu>
  </body>
</tmx>
```

Example 1. A TMX document.

The corresponding representation in MLIF is as follows:

```
<MLDC>
 <GI>
  <sourceFormat>TMX</sourceFormat>
  <formatVersion>1.4</formatVersion>
  <creationDate>20040731T164933Z</creationDate>
  <creationTool>Heartsome TM Server</creationTool>
  <creationToolVersion>1.0.1</creationToolVersion>
 </GI>
 <GroupC>
  <MultiC>
   <creationIdentifier>1091303313515</creationIdentifier>
   <creationDate>20020930T004233Z</creationDate>
   <MonoC xml:lang="fra">
    <SegC>Le processus de contrôle de qualité en dix étapes
          qu'il a créé il y a plus de 1300 ans est beaucoup
          plus complet et précis que ceux existant
          aujourd'hui.</SegC>
   </MonoC>
   <MonoC xml:lang="eng">
    <SegC>His 10-stage quality control process initiated
          more than 1300 years ago is far more thorough and
          exacting than any existing today.</SegC>
   </MonoC>
  </MultiC>
 </GroupC>
```

Example 2. The MLIF version of Example 1.

As we can see, TMX is nearly isomorphic to the MLIF metamodel. The core elements of the TMX macro-structure map to MLIF components as follows:

- <tmx> matches onto the <MLDC> component.
- <header> maps onto the <GI> component.
- <body> is a container for <tuv> element and does not map onto any component of the <MLIF> metamodel.
- <tu> maps onto the <MultiC> component.
- <tuv> maps onto the <MonoC> component.
- <seg> maps onto the <SegC> component.

3.1 Computer-aided translation use case

MLIF promotes the use of linguistic information in computer-aided translation tools based on translation memory in order to allow these tools to produce translation for new words and new sentences which are not in the translation database.

For example, with a translation memory containing the English sentence "The meal is nice." and its translation in French "Le repas est bon.", current CAT tools such as SDL TRADOS Translator's Workbench, DejaVu and Star Transit are not able to propose the expected translation for the input: "The meals are nice." although the word lemmas of "The meal is nice" and "The meals are nice" are matching. This weakness is due to the fact that these tools use limited linguistic criteria during the translation process.

```
<tmx version="1.4">
```

```
 <header
   creationtool="TRADOS Translator's Workbench for Windows"
   creationtoolversion="Edition 7 Build 719"
   segtype="sentence"
   o-tmf="TW4Win 2.0 Format"
   adminlang="EN"
   srclang="EN"
   datatype="rtf"
   creationdate="20090922T140557Z"
   creationid="SEMMAR"/>
 <body>
  <tu creationdate="20090922T140653Z" creationid="SEMMAR">
   <tuv xml:lang="EN">
    <seg>The meal is nice.</seg>
   </tuv>
   <tuv xml:lang="FR">
    <seg>Le repas est bon.</seg>
   </tuv>
  </tu>
 </body>
</tmx>
```

Example 3. Data produced by TRADOS Translator's Workbench 7.

In order to translate the sentence "The meals are nice.", an MLIF-compliant tool may implement the following procedure:

Step-1: Add linguistic properties to all the words within the translation memory.

```
<MultiC>
<creationIdentifier>SEMMAR</creationIdentifier>
<creationDate>20090922T140653Z</creationDate>
 <MonoC xml:lang="eng">
  <SegC>The meal is nice.</SegC>
 </MonoC>
 <MonoC xml:lang="fra">
  <SegC>Le repas est bon.</SegC>
 </MonoC>
</MultiC>
<MultiC class="translation">
 <MonoC xml:lang="eng">
  <SegC class="word" lemma="the" pos="definiteArticle">The</SegC>
  <SegC
    class="word"
    lemma="meal"
    pos="commonNoun"
    mfeature="singular">meal</SegC>
  <SegC
    class="word"
    lemma="be"
    pos="verb"
    mfeature="present thirdPerson singular">is</SegC>
  <SegC class="word" lemma="nice" pos="qualifierAdjective">nice</SegC>
  <SegC class="word" lemma="." pos="mainPunctuation">.</SegC>
 </MonoC>
<MonoC xml:lang="fra">
  <SegC
    class="word"
    lemma="le"
    pos="definiteArticle"
    mfeature="masculine singular">Le</SegC>
  <SegC
    class="word"
    lemma="repas"
```

```
    pos="commonNoun"
    mfeature="masculine singular">repas</SegC>
  <SegC
    class="word"
    lemma="être"
    pos="verb"
    mfeature="present thirdPerson singular">est</SegC>
  <SegC
    class="word"
    lemma="bon"
    pos="qualifierAdjective"
    mfeature="masculine singular">bon</SegC>
  <SegC class="word" lemma="." pos="mainPunctuation">.</SegC>
 </MonoC>
</MultiC>
```

Example 4. The MLIF version of Example 3.

Step-2: Run a part-of-speech tagger on the sentence in order to obtain the right morpho-syntactic word categories.

Step-3: Translate the lemmas by the means of an English to French bilingual lexicon.

Step-4: Lookup in a French lexicon of inflected forms in order to retrieve the right inflected form by means of the lemma and the morphological features.

Step-5: Generate the translation of "The meals are nice." by substitution of each English word by its French inflected form as follows: The sentence "The meals are nice." is translated with the sentence "Les repas sont bons."

3. Multimedia and Multilingual Textual Information use case

SmilText is a module defined in the context of SMIL 3.0 W3C recommendation. It has the potential to be an important application context for MLIF as it associates and synchronizes multimedia and textual content.

General timing mechanisms from the SMIL (Synchronized Multimedia Integration Language) recommendation may be used in MLIF compliant content to provide synchronization mechanisms of textual content. The following attributes are thus integrated in the overall MLIF specification: begin, end, dur (short for duration).

The basic use case for articulating MLIF and SMIL consists in producing a monolingual SMIL output out of a multilingual representation expressed in an MLIF compliant format. This results from the selection of the content corresponding to a chosen language and its integration into one or several <smilText> containers, for instance embedded in a <seq> construct. When applicable the existing timing information are propagated into the SMIL representation.

In this context, the core mappings between MLIF and the SMIL Text specification would be the following ones:

- <MonoC> components should map onto <smilText> elements, together with the corresponding attributes (in particular, language).
- <SegC> components should map unambiguously onto <tev> elements, together with the corresponding descriptors (in particular temporal ones).

The actual embedding of multilingual content within a single SMIL representation should be based on the a <switch> constructs within the following skeleton:

```
<switch>
 <par systemLanguage="en">
  <smilText
    xml:id="TE30"
    region="Contents"
    dur="12s"
    its:dir="ltr"
    xml:lang="eng"
    its:translate="yes"> ...
  </smilText>
 </par>
 <par systemLanguage="fr">
  <smilText
```

```
    xml:id="TF30"
    region="Contents"
    dur="12s"
    its:dir="ltr"
    xml:lang="fra"
    its:translate="yes"> ...
  </smilText>
 </par>
</switch>
```

Example 5. Skeleton of multilingual content within a single SMIL file.

Other non-temporal attributes such as region are not covered by the MLIF specification and should thus be created independently from the MLIF compliant structure. This equivalence can be used to conversely generate MLIF compliant content from a SMIL representation. The associated use case is typically the preparation of an MLIF compliant structure that will later contain further translation(s).

The core elements of smilText map to MLIF components as follows:

- The <smilText> element functions as a logical and temporal structuring element that allows the inclusion of in-line text content into a SMIL presentation. SmilText can also be used as an external, stand-alone timed text format. This is accomplished by using the SMIL 3.0 SmilText profile.
- The <tev> element defines a "temporal moment" within a block of smilText content. Depending on the values of the begin or next attributes, it determines a scheduling time at which the associated text content (up to the following tev or clear element or the end of the smilTextelement) is rendered.
- Mapping on SegC, the <clear> element defines a "temporal moment" within a block of smilText content at which the full contents of the rendering area are cleared. Depending on the values of the begin or next attributes, it also determines a scheduling time at which the associated text content (up to the following tev or clear element or the end of the smilText element) is rendered. This element is functionally equivalent to the tev element, except that it has a side-effect of clearing the rendering area before any new content is rendered.

Besides, the following SMIL attributes map as follows:

- “dur” maps onto the <duration> data category.
- “begin” maps onto the <begin> data category.
- “next” maps onto the <next> data category.

4. Conclusion

With this brief presentation of MLIF, we have showed that MLIF provides a generic specification to represent and manage multilingual information and, thus to assure interoperability with existing standards in the domains of translation and localization. With the help of appropriate software tools, MLIF is an important step forward the representation of independent (but incomplete) structures within a coherent and normalized framework.

In order to validate our proposals, we are currently using MLIF within the following projects:

1. The ANR WebCrossling project that aims to develop a machine translation tool based on a cross-language information retrieval approach. This tool uses translation memories enriched with linguistic information and validated by professional translators via a Web 2.0 application.

2. The ITEA2 SEMbySEM project that aims at defining tools and standards for the supervision and management of complex and dynamic systems by using a semantic abstract representation of the system to be supervised or managed. As we want the system to conform to an end-user’s point of view, the conceptual information must be available and presentable in the end-user’s language (http://www.sembysem.org).

3. The ITEA2 METAVERSE1 project that will provide a standardized global framework that enables the interoperability between virtual worlds (as for example Second Life, World of Warcraft, IMVU, Google Earth and many others) and the real world (sensors, actuators, vision and rendering, social and welfare systems, banking, insurance, travel, real estate and many others). The ‘Metaverse for all’ will be a special attention point aiming at the eInclusion, of minorities in the society (http://www.metaverse1.org).

5. References

Cruz-Lara S., Bellalem N., Ducret J., and Krammer I. (2008). Standardising the Management and the Representation of Multilingual Data: the Multi Lingual Information Framework. *Topics in Language Resources for Translation and Localisation*. Editor Elia Yuste, pp. 151-172. John Benjamins Publishers.