Some general remarks about Timed Text (TT) Authoring Format 1.0 – Distribution Format Exchange Profile (DFXP)

http://www.w3.org/TR/2006/WD-ttaf1-dfxp-20060427/

Samuel CRUZ-LARA
LORIA / INRIA Lorraine
Samuel.Cruz-Lara@inria.fr

Introduction.

This short document presents some general remarks about TimedText DFXP. First, we will present some useful information about the "Multi Lingual Information Framework" (MLIF) [1]; second, we will discuss briefly about TimedText DFXP in the framework of multiple natural languages, natural language granularity and natural language coverage. Finally, we will present a short conclusion and some final remarks.

The "Multi Lingual Information Framework".

MLIF is an ISO's "New Work Item Proposal" (NWIP) that has been recently sent to ISO's TC37 / SC4 "Linguistic Resources Management". This NWIP has motivated a classical 3 months ballot process and we are expecting to have the final result at the end of August 2006. If the result of this ballot process is positive, MLIF will become an ISO's Working Draft, and hopefully a little later, an ISO's standard.
Linguistic information plays an essential role in the management of multimedia information, as it bears most of the descriptive content associated with more visual information. Depending on the context, it may be seen as the primary content (text illustrated by pictures or videos), as documentary content for multimedia information, or as one among several possible information components in specific contexts such as interactive multimedia applications.
Linguistic information can appear in various formats: spoken data in an audio or video sequence, implicit data appearing on an image (caption, tags, etc.) or textual information that may be further presented to the user graphically or via a text to speech processor.

In this context, dealing with multilingual information is crucial to adapting the content to specific user targets. It requires one to consider potential situations where the linguistic information contained in a multimedia sequence is either already conceived in such way that it can be adapted on the fly to the linguistic needs of user, or by using an additional process where content should be adapted before presenting it to the user.

MLIF aims at proposing a specification platform for a computer-oriented representation of multilingual data within a large variety of applications such as translation memories, localization, computer-aided translation, multimedia, or electronic document management. As with the "Terminological Markup Framework", used in terminology [ISO 16642], the MLIF will introduce a metamodel in combination with chosen data categories that will be integrated within the TC37 data category registry in order to allow the description of any specific

[1] The "Multi Lingual Information Framework" (MLIF). New Work Item Proposal.

domain. The standard will thus provide a way to validate any instance of this metamodel, as well as, interoperability principles with numerous translation and localization standards.

The extremely fast evolution of the technological development in the sector of Communication and Information Technologies, and in particular, in the field of natural language processing, makes particularly acute the question of standardization. The issues related to this standardization are of an industrial, economic and cultural nature. The control of the interoperability between the industrial standards currently used for localization (XLIFF), translation memory (TMX) , or any other Multi Lingual Markup Language (ML2), constitutes a major objective for a coherent and global management of these data. The MLIF could be associated to several multimedia standards such as MPEG-4 [ISO/IEC 14496] and MPEG-7 [ISO/IEC 15938], and W3C SMIL, in order to handle multilingual data within several multimedia applications such as, interactive TV, video conferencing, subtitling, karaoke and accessibility. The MLIF may also be used in cultural heritage related activities such as, digital museums, e-learning and electronic document management.
As with the "Terminological Markup Framework" (TMF), used in terminology, the MLIF will introduce a metamodel in combination with chosen data categories. These data categories will be derived as a subset of a Data Category Registry (DCR) [rev ISO 12620], in order to ensure interoperability between several multilingual applications and corpora. A Data Category Specification (DCS) will define, in combination with the metamodel, the various constraints that apply to a given domain-specific information structure or interchange format. A DCS and a metamodel represent the organization of an individual application and the organization of a specific domain.
The MLIF should be considered as a unified conceptual representation of multilingual content. It is not meant to replace or to compete with any other existing standard. Rather, the MLIF is being designed with the objective of providing a common conceptual model and a platform allowing interoperability among several translation and localization standards, and by extension, their committed tools. The asset of MLIF is the interoperability which allows experts to gather, under the same conceptual unit, various tools and representations related to multilingual data. In addition, MLIF will also make it possible to evaluate and to compare these multilingual resources and tools.
The description of all different XML elements will be done by using RelaxNG [ISO 19757-2] with the help of ODD , which is the creation and documentation language for XML schemas proposed by the TEI (Text Encoding Initiative). This follows a recent decision taken by the World Wide Web Consortium: The Internationalization Tag Set (ITS) Version 1.0 (http://www.w3.org/TR/its/) is the very first W3C specification written using the TEI's ODD language for creating and documenting XML schemas.

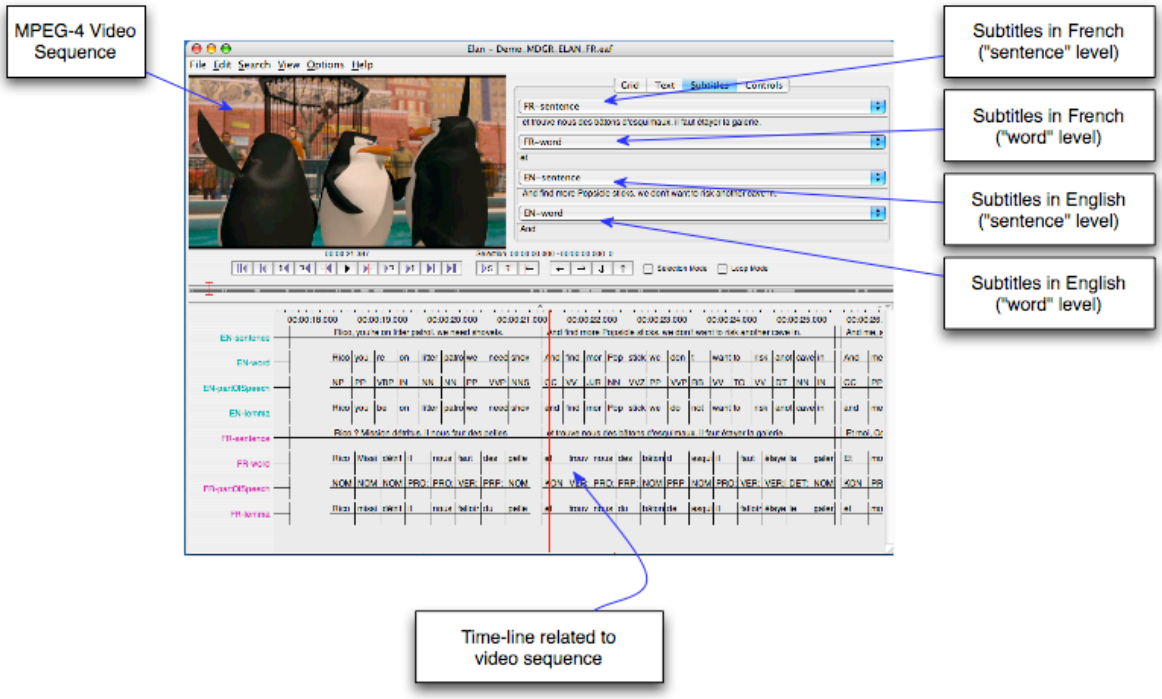
TimedText DFXP and Natural Language.

First of all, one should note that our knowledge about TimedText DFXP is rather incomplete. We have only consulted some documents published by the W3C as "Timed Text (TT) Authoring Format 1.0 – Distribution Format Exchange Profile (DFXP)" and "Timed Text (TT) Authoring Format 1.0 Use Cases and Requirements".

In the second document (Timed Text (TT) Authoring Format 1.0 Use Cases and Requirements), we can see that multiple natural languages (R201), natural language granularity (R203), and natural language coverage (R202) are mentioned as TimedText requirements. However, in the framework of natural language granularity, for example, no

information is given in order to understand how natural language granularity is taken into account. The only information we have is that natural language will be identified by using the *xml:lang* attribute.
In our opinion, considering a specific granularity of segmentation and description (i.e., morphological description, syntactical annotation, terminological description, etc) allow to effectively describe elementary linguistic segments (i.e. sentence, syntactical component, word, syllables, part of speech, etc). This is a very important issue in the framework of associating textual information to multimedia presentations. For example, being able to deal with elementary linguistic segments is essential for highlighting some specific words in a subtitle or a in a caption. In addition, for some practical reasons related to language learning, for example, it should be interesting for a multimedia presentation to be able to deal with multilingual alignment. This multilingual alignment needs a good description and handling of natural language granularity.

The following figure shows several levels of granularity associated to a multimedia MPEG-4 presentation.



This demo has been prepared in the framework of ITEA's Passepartout project by using the following tools:

1. Tree Tagger (University of Stuttgart D)
    - http://www.ims.unistuttgart.de/projekte/corplex/TreeTagger/)
2. ELAN (Max Planck Institute, Nijmegen NL)
    - http://www.mpi.nl/tools/elan.html

Conclusion and final remarks.

This short document has presented some general remarks about TimedText DFXP. Our knowledge of TimedText, as we have indicated in this document, is rather incomplete. However, having consulted some documents that the W3C has recently published, we think

that some important aspects related to natural language, in particular those related to a specific granularity needed to effectively describe elementary linguistic segments, is only partially taken into account by TimedText DFXP.
In addition, and this is mainly a consequence of the precedent consideration, defining linguistic alignments between two or more different natural languages, seems to be impossible. So, multilinguality is also only partially taken into account by TimedText DFXP.
If, in the framework of associating textual information to multimedia presentations, timing is a very important issue, being able to deal with several levels (i.e. granularity) of linguistic segments is also a very important matter.
If for a majority of multimedia presentations a granularity of “primary text” may be enough, for other applications (i.e. universal access, language learning, karaoke, etc) a more detailed description of linguistic segments would be needed.