HAL
archives-ouvertes.fr

Interoperability between translation memories and localization tools by using the MultiLingual Information Framework

Samuel Cruz-Lara, Nadia Bellalem, Julien Ducret, Isabelle Kramer

▶ **To cite this version:**

Samuel Cruz-Lara, Nadia Bellalem, Julien Ducret, Isabelle Kramer. Interoperability between translation memories and localization tools by using the MultiLingual Information Framework. European Association for Machine Translation - EAMT 2006, Jun 2006, Oslo/Norway, 2006. <inria-00105651>

HAL Id: inria-00105651
https://hal.inria.fr/inria-00105651

Submitted on 11 Oct 2006

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Interoperability between translation memories and localization tools by using the MultiLingual Information Framework

Samuel Cruz-Lara, Nadia Bellalem, Julien Ducret, Isabelle Kramer

LORIA / INRIA Lorraine
Campus Scientifique - BP 239
54506 Vandoeuvre-lès-Nancy, France
{Samuel.Cruz-Lara, Nadia.Bellalem, Julien.Ducret, Isabelle.Kramer}@loria.fr

Abstract. The scope of research and development in the localization and translation memory process development is huge. Several formats have been developed of specific interest for localization and translation such as XLIFF and TMX. The associated software industry has thus developed several well-known tools committed to these formats: TRADOS, SDLX, DEJAVU, etc. When we closely examine these formats, we find that they have many overlapping features. They work well in the specific field they are designed for, but they lack the synergy that would make them interoperable when using one type of information in a slightly different context. The Multi Lingual Information Framework (MLIF) is being designed with the objective of providing a common conceptual model and a platform allowing interoperability among several translation and localization formats, and by extension, their committed tools. MLIF does not have the role to substitute or compete with existing standards: MLIF should be considered as a common abstract high-level framework in which the overlapping features of several existing formats may be handled independently and separately. MLIF would save time and energy for different translation and localization groups and would provide synergy to work in collaboration. MLIF is a way of opening the field of localization and translation at other communities (the multimedia community, for example) and, a way of finding there, new outlets or actors, sources of innovation.

1. Introduction

Standards make an enormous contribution to most aspects of our lives. People are usually unaware of the role played by standards in raising levels of quality, safety, reliability, efficiency and interoperability - as well as in providing such benefits at an economical cost. The scope of research and development in localization and translation memory process development is very large, many industrial standards and their associated software industry have been developed, for example, SDLX for XLIFF [1] and, TRADOS and Déjà Vu for TMX [2]. The current versions of translation tools on the market work quite well, but previous versions sometimes created their own "flavor" of TMX or XLIFF which could not readily be imported by other tools, so export files were to be changed before an import.

Of course, these standards were developed for make possible the exchange of data between tools. The question is, how well can the data that has been exchanged can be used. Modeling corresponds to the need to describe and compare existing interchange formats in terms of their informational coverage and the conditions of interoperability between these formats and hence the source data generated in them. One of the issues here is to explain how an uniform way of documenting such databases considering the heterogeneity of both, their formats and their descriptors. We also seek to answer the demand for more flexibility in the definition of interchange formats without any change for the tools. Such an attempt should lead to more general principles and methods for

analyzing existing multilingual databases and mapping them onto any chosen multilingual interchange format.

2. Introduction to TM tools

2.1. Cycle of life of multilingual information

A multilingual software product should aim at supporting document indexing, automatic and/or manual computer-aided translation, information retrieval, subtitle handling for multimedia documents, etc. Dealing with multilingual data is a three steps process: production, maintenance (updated, validation, correction) and consumption (use). For example, depending of the tools, that produced the TMX file, it can be bilingual or multilingual. When we import a multilingual TMX file into a bilingual project (e.g. TMX to XLIFF file), we will only import the relevant languages. If we don't have a common format, some maintenance problems can appear as well as lack of synergy and several overlapping issues. Multilingual data are not only used in the framework of translation and localization, and they also belong to terminology, index system, e-learning, etc. Each specific domain can improve the quality of information of each other. For example, linguistic information (e.g. part of speech, lemma, etc) could be added to multilingual data, in order to expand the translation memory process.

2.2. List of TM tools

In this part we will discuss about two major problems of dealing with different tools and different formats: formatting and segmentation. Although TM Tools are based on the same basic idea, we must note that for the same sentence each tool proposes rather different ways to implement the required formatting information: on the one hand, formatting is applied to the source and target texts of a translation unit and this formatting is not exported to the corresponding TMX file; on the other hand, formatting is sometimes exported to the TMX file. In the following table (see Figure 1), the sample sentence "the sentence contains different formatting information" is represented in TMX by using several tools [3]. Some of these tools use external files to store formatting information (Déjà Vu, SDLX), but all of them use different ways of encoding that information.

TRADOS 6.5	DÉJÀ VU	SDLX
<seg> This <ut>{\b /ut>sentence<ut>}</ut> contains <ut>{\i </ut>different<ut>}</ut> <ut>{\ul </ut>formatting information<ut>}</ut>. </seg>	<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>	<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>

Figure 1. Comparison of tools formatting

In addition, the segmentation rules used by TM tools are not compatible: each tool applies his own rule to split the text into various segments. In a same sentence some tools consider various separators. For example the semi-colon is considered as a separator for Déjà Vu, but not for SDLX.

Segmentation organizes and structures the data. If every one uses his own rules, the exchange is no more possible; that's why SRX [4] for several years tries to normalize segmentation rules. SRX guidelines are useful to evaluate translation memory qualities and ensure interoperability of multilingual data.

2.3. High-level Representation and Interoperability

One may think that, as a TM is really specific of a kind of translation job, transforming a TM from one format to another is useful only when a client switches from one translation tool or provider to another. In the reality, this would almost never been necessary.

However, as we shall explain in the following sections, the main objective of MLIF is not really to facilitate transformations from one format to another, but well beyond that, to be able to represent multilingual data in the most independent possible manner (by using an abstract high-level representation) with respect to any specific format.

In the following sections, we shall describe how MLIF is being designed and how we can use it. By now, it is very important to understand that if we have previously used an example based on formatting issues (see Figure 1), MLIF is being designed to be used in a much more general way.

3. Terminology of normalization

In the same way as "Terminological Markup Framework" (TMF) [5] in terminology, MLIF will introduce a structural skeleton (metamodel) in combination with chosen data categories [6], as a means of ensuring interoperability between several multilingual applications and corpora.

3.1. Metamodel

A metamodel does not describe one specific format, but acts as a kind of high level mechanism based on the following elementary notions: structure, information, and methodology. The structuring elements of the metamodel are called "components" and they may be "decorated" with information units. A metamodel should also comprise a flexible specification platform for elementary units. This specification platform should be coupled to a reference set of descriptors that should be used to parameterize specific applications dealing with content.

3.2. Data Categories

A metamodel contains several information units related to a given format, which we refer to as "Data Categories". A selection of data categories can be derived as a subset of a Data Category Registry (DCR) ensuring that the semantic of these data categories is well defined and accepted by an ISO committee. A data category is the generic term that references a concept. There is one and only one identifier for a data category in a DCR. All data categories are represented by a unique set of descriptors. For example, the data category /primaryText/ indicates a linguistic material which is the object of study. A Data category Selection (DCS) is needed in order to define, in combination with a metamodel, the various constraints that apply to a given domain-specific information structure or interchange format. A DCS and a metamodel can represent: the organization of an individual application, or the organization of a specific domain.

3.3. Implementation

The means to actually implement a standard is to instantiate the metamodel in combination with the selection of data categories. This includes mappings between data categories and vocabularies used to express them (e.g. as an XML element or a database field). A DCS is firstly used to specify constraints on the implementation of a metamodel instantiation, and secondly to provide the necessary information for implementing filters that convert one instantiation to another and allows to produce a "Generic Mapping Tool" (GMT) representation. The architecture of the metamodel, whatever the standard we want to specify, remains unchanged. What is variable are the data categories selected for a specific application. Indeed, the metamodel can be considered in an atomic way, in the sense that starting from a stable core, a multitude of data can be worked out for plural activities and needs.

4. MLIF

Linguistic structures exist in a wide variety of formats ranging from highly organized data (e.g. translation memory) to loosely structured information. The representation of multilingual data is based on the expression of multiple views representing various levels of linguistic information, usually pointing to primary data (e.g. part of speech tagging) and sometimes to one another (e.g. reference annotation based on basic phrase structure annotation). The following model identifies a class of document structures, which could be used to cover a wide range of multilingual formats, and provides a framework, which can be applied using XML. MLIF is being designed in order to provide a generic structure that can establish basic foundation for all these standards.

4.1. MLIF Metamodel

A MLIF document has a hierarchical structure as shown in Figure 1. This document will have "MultilingualDataCollection" as the root level element, which content two major components: the "GlobalInformation" element and the "MultiLingualComponent" element.

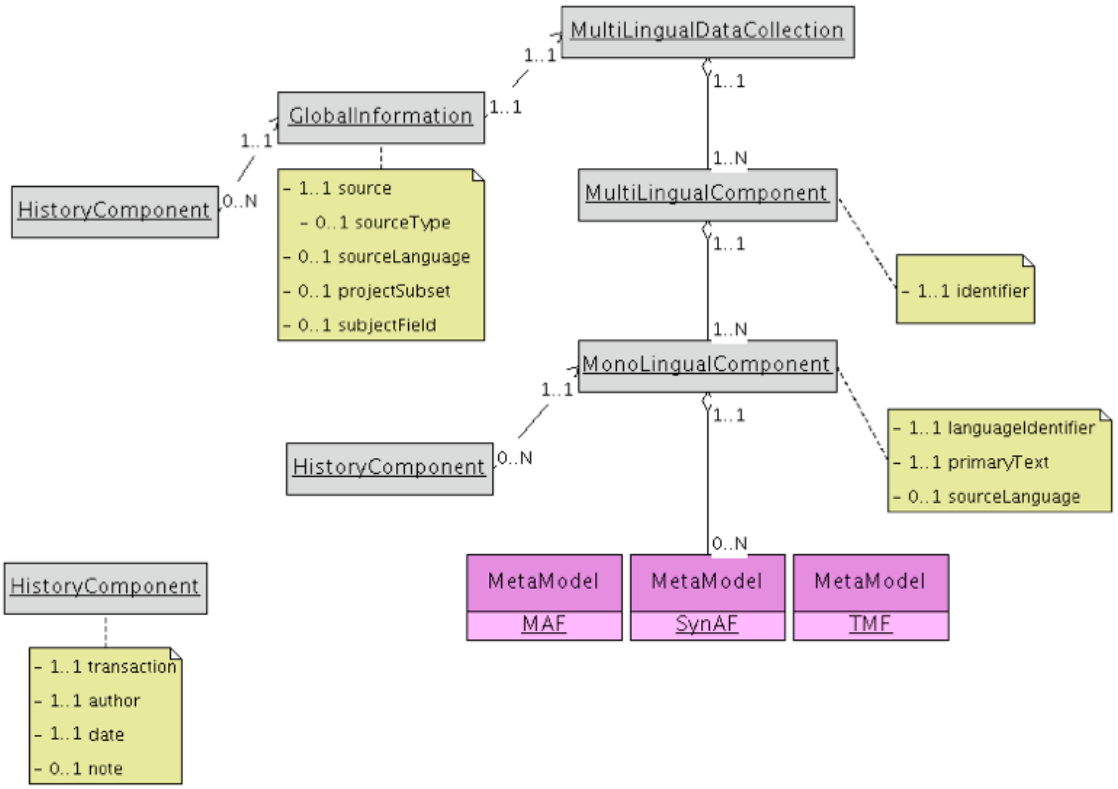

Figure 2. MLIF Metamodel and related Data Categories

The "GlobalInformation" element can be considered as a header element which contents metadata related to the document as source of the document and other administrative information. In a document we can have one or more multilingual components. A "MultiLingualComponent" contains information that belongs to the linguistic unit (e.g. a single sentence or a paragraph, etc), descriptive informations (e.g. domain of application) or administrative datas (e.g. transaction, identifier, alias). Each "MultiLingualComponent" must content one or more "MonoLingualComponent" elements. A "MonoLingualComponent" is the linguistic unit in a given language. It could be a source text or a translation of this text into another language. The "HistoryComponent" is a generic component allowing to trace modifications on the component it is anchored to (e.g., creation, modification, validation). It can be anchored onto any component of the metamodel. In MLIF metamodel, the "HistoryComponent" may be anchored to the "GlobalInformation"

component or to the "MonoLingual Component". In the "GlobalInformation" component, it keeps all information related to any modification on the context or on the domain; in the "MonoLingualComponent", it allows keeping all evolutions or any enhancement of the content.

It should be noted that in order to provide a larger description of the linguistic content, MLIF metamodel (see Figure 2) allows anchoring of other metamodels, such as MAF (Morphological Description), SynAF (Syntactical Annotation), TMF (Terminological Description), or any other metamodel based on ISO 12620:2003.

For understanding what is MLIF, it is important to distinguish what depends, on the one hand, on the metamodel or, on the other hand, on the data categories. In fact, each structural node can be qualified by a group of basic or compound information units. A basic information unit describes a property that can be directly expressed by means of a data category. A compound information unit corresponds to the grouping at one level of several basic information units, which taken together, express a coherent unit of information.

4.2. Some Possible Data Categories for MLIF

Global Information

/source/

- A complete citation of the bibliographic information pertaining to a document or other resource.
- Reference to a resource from which the present resource is derived.

/sourceType/

- In multilingual and translation-oriented language resource or terminology management, the kind of text used to document the selection of lexical or terminological, equivalents, collocations, and the like.

/sourceLanguage/

- In a translation-oriented language resource or terminology database, the language that is taken as the language in which the original text is written.
    - Both parallel and background texts serve as sources for information used in documenting multilingual terminology entries

/projectSubset/

- An identifier assigned to a specific project indicating that it is associated with a term, record or entry.

/subjectField/

- A field of special knowledge.

Multilingual Component

/identifier/

- A unique name.
    - Dublin Core equivalent: DC:Identifier

Monolingual Component

/languageIdentifier/

- A unique identifier in a language resource entry that indicates the name of a language.

/primaryText/

- Linguistic material which is the object of study.

/sourceLanguage/

- In a translation-oriented language resource or terminology database, the language that is taken as the language in which the original text is written.
    - The identifiers specified in ISO 639 should be used:
        - en = English
        - fr = French
        - es = Spanish (Español)
        - de = German (Deutsch)
        - ru = Russian
        - …

4.3. Introduction to GMT

GMT can be considered as a XML canonical representation of the generic model. The hierarchical organization of the metamodel and the qualification of each structural level can be realized in XML by instantiating the abstract structure shown above (Figure 2) and associating information units to this structure. The metamodel can be represented by means of

a generic element <struct> (for structure) which can recursively express the embedding of the various representation levels of a MLIF instance. Each structural node in the metamodel shall be identified by means of a type attribute associated with the <struct> element. The possible values of the type attribute shall be the identifiers of the levels in the metamodel:

- MultilingualDataCollection;
- GlobalInformation;
- MultiLingualComponent;
- MonoLingualCompon.

Basic information units associated with a structural skeleton can be represented using the <feat> (for feature) element. Compound information units can be represented using the <brack> (for bracket) element, which can itself contain a <feat> element followed by any combination of <feat> elements and <brack> elements. Each information unit must be qualified with a type attribute, which shall take as its value the name of a standardized data category or one user-defined data category.

```
<tu creationdate="20060128T133704Z" creationid="MLIFTeam">
    <tuv lang="en">
          <seg>This is the first sentence.</seg>
    </tuv>
    <tuv lang="de">
          <seg>Dies ist der erste Satz.</seg>
    </tuv>
</tu>
```

Figure 3. A TMX Example

In Figure 3, we found two strong structural elements in TMX : the <tu> element and a <tuv> element. These two TMX elements will correspond to the following MLIF structurals elements: <tu> corresponds to "MultiLingualComponent" and <tuv> corresponds to "MonoLingualComponent".

```
<struct type="MultilingualDataCollection">
    <struct type="GlobalInformation">
          <feat type="source">TMX Example</feat>
          <struct type="HistoryComponent">
                <feat type="transaction">creation</feat>
                <feat type="date">20060128T133704Z</feat>
                <feat type="author">MLIFTeam</feat>
          </struct>
    </struct>
    <struct type="MultiLingualComponent">
          <feat type="identifier">503</feat>
          <struct type="MonolingualComponent">
                <feat type="languageIdentifier">en</feat>
                <feat type="primaryText">This is the first sentence.</feat>
          </struct>
          <struct type="MonolingualComponent">
                <feat type="languageIdentifier">de</feat>
                <feat type="primaryText"> Dies ist der erste Satz.</feat>
          </struct>
    </struct>
</struct>
```

Figure 4. MLIF implementation

4.4. TMX and MLIF interaction

Figure 5 (see below) illustrates the interaction between TMX and MLIF. This diagram includes the following steps: extraction, translation, merging. The starting point is a TMX document which linguistic content is in English (EN) and in German (DE). The extraction process (1) allows to obtain in one side a "Skeleton File" (2)which contains all TM formatting information and in an another part a MLIF file (3) in which only relevant linguistic information is stored. As most translators (human or automatic) work with TMX software oriented-tools, a XSL style-sheet allows to transform a MLIF document into a TMX document. This file does not contain any formatting information. Once the translator (human or automatic) has added the related Japanese translation, another XSL style-sheet allows to transform a TMX document into a MLIF document (4). Finally, the new MLIF document (this containing the Japanese translation) is merged with the "Skeleton File" in order to obtain a new TMX formatted document (5).

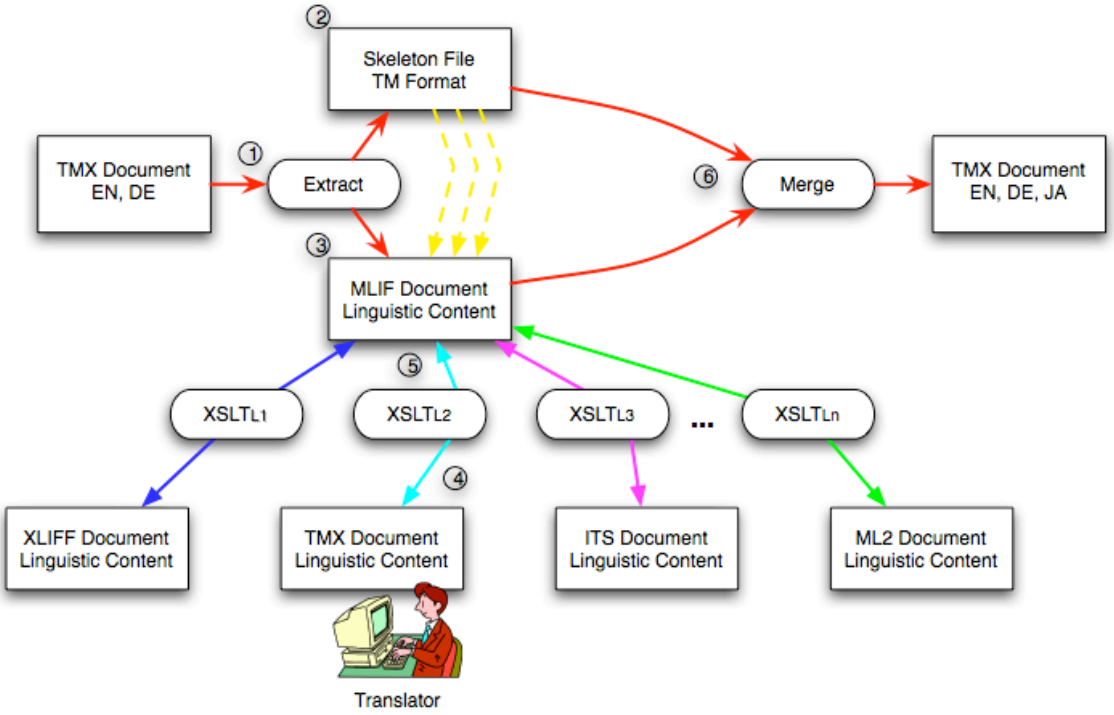

Figure 5. TMX and MLIF interaction

One should note that the asset of MLIF is the interoperability that allows experts to gather, under the same conceptual unit, various tools and representations related to multilingual data. So, the presence of XLIFF and ITS in Figure 5 means that, by using MLIF, the interoperability between XLIFF, TMX, and ITS may become possible.

It is important to recall that MLIF does not have the role to substitute or to compete with any existing standard. 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.

5. Conclusion

We have presented MLIF (Multi Lingual Information Framework): a high-level model for describing multilingual data. MLIF can be used in a wide range of possible applications in the translation/localization process in several domains. This paper should be considered as a first step towards the definition of abstract structures for the description of multilingual data. The idea in a near future is to be able to implement interoperable software libraries

which can be independent of the handled formats. A first "informal" presentation of MLIF at AFNOR (Association Française pour la Normalisation - ISO's French National Body) on December 7th, 2005. We have obtained several very positive comments about our draft proposal. It should also be noted that a "new work item proposal" (nwip) has been recently sent to ISO TC37 / SC4 subcommittee: a ballot process has been started. If the result of this ballot process is successful, MLIF will officially become an ISO's Working Draft (WD).

In addition, within the framework of ITEA "Passepartout" project [7], we are experimenting with some basic scenarios where MLIF is associated to XMT (eXtended MPEG-4 Textual format [8]) and to SMIL (Synchronized Multimedia Integration Language [9]). Our main objective in this project is to associate MLIF to multimedia standards [10], [11], [12] (e.g. MPEG-4, MPEG-7, and SMIL) in order to be able, within multimedia products, to represent and to handle multilingual content (subtitles, retrieval of textual information by user interaction, …) in an efficient, rigorous and interactive manner.

6. References

[1] XLIFF. Oasis (2003). XML Localisation Interchange File Format, http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=xliff. Accessed 2003-10-31.

[2] TMX. Oscar / Lisa (2000) Translation Memory eXchange, http://www.lisa.org/tmx. Accessed 2001-06-27.

[3] TMX and SRX Exchanging TM Data. Angelika Zerfass, Consultant and Trainer for Translation Tools. LRC-X Conference, University Of Limerick, Ireland. 13-14 September 2005.

[4] SRX. Segmentation Rules eXchange. SRX 1.0 Specification. Oscar Recommendation 20 April, 2004. http://www.lisa.org/standards/srx/srx.html .

[5] TMF. ISO 16642 (2003) Computer applications in terminology -- Terminological markup framework, Genèva, International Organization for Standardization.

[6] ISO 12620 (1999) : Computer applications in terminology -- Data categories,

[7] ITEA "Information Technology for European Advancement". Passepartout project "Exploitation of advanced AV content protocols (MPEG 4/7)" ITEA 04017.

http://www.itea2.org/public/project_leaflets/PASSEPARTOUT_profile_oct-05.pdf .

[8] XMT. extended MPEG-4 Textual format. ISO/IEC FCD 14496-11, Information technology -- Coding of audio-visual objects -- Part 11: Scene description and application engine; ISO/IEC 14496-11/Amd 4, XMT & MPEG-J extensions.

[9] Synchronized Multimedia Integration Language (SMIL 2.0). World Wide Web Consortium. http://www.w3.org/TR/smil20/ .

[10] S. Cruz-Lara, S. Gupta, & L. Romary (2004). Handling Multilingual content in digital media: The Multilingual Information Framework. EWIMT-2004 European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology. London, UK. November 2004.

[11] S. Cruz-Lara, S. Gupta, J.D. Fernández García and L. Romary (2005). Multilingual Information Framework for Handling Textual Data in Digital Media. IEEE AMT 2005, The Third International Conference on Active Media Technology. Takamatsu, Kagawa, Japan. May 2005.

[12] S. Cruz-Lara, N. Bellalem, J. Ducret and I. Kramer. Interactive Handling of Multilingual Content within Digital Media. EuroiTV 2006 Beyond Usability, Broadcast, and TV. Workshop "Present and Future of Software Graphics Architectures for Interactive Television". Athens, Greece. May 2006.