LREC 2008 Tutorial

LISA standards and guidelines for multilingual text processing and management

Content

Exchange standards play an important role in today's localization and translation industry. By providing agreed upon ways of representing linguistic data they allow various applications to interoperate and share data. The Localization Industry Standards Association (LISA, http://www.lisa.org) has been particularly active in development of standards for linguistic data. This tutorial will describe the various LISA standards and specifications and their relationship to other standards initiatives:

- **Translation Memory eXchange** (TMX) is the industry standard for the exchange of aligned translation memory databases. LISA research has shown that some organizations have millions of dollars invested in translation memory data. This investment, coupled with ongoing consolidation and changes in the lineup of TM vendors and products, makes the ability to reuse TM data and exchange TM databases of vital importance if these assets are to maintain their value.

- **Term-Base eXchange** (TBX) is the industry standard for representing terminology databases in XML format. Terminology is a key component to achieving quality translations (LISA research has demonstrated that terminology errors result in negative impression of product quality on par with seemingly more serious functional errors), yet many organizations do not manage terminology or use simple tools like Microsoft Excel to store their terminology. TBX, which is now a joint work item between LISA and ISO (ISO 30042) simplifies the use and dissemination of terminological data and helps organizations leverage these assets during the translation process.

- **Segmentation Rules eXchange** (SRX) provides a regular expression-based XML formalism to describe how text is split into "segments" (such as sentences or paragraphs). When coupled with TMX, it allows linguistic tools to describe how they segmented text during the translation process so that other tools can emulate this behavior in order to better use text produced by them. It also has potential use in any field that requires text segmentation. Planned submission to ISO will focus on allowing national bodies to define default segmentation rules sets for their languages in order to better facilitate use of linguistic tools with those languages.

- **xml text memory** (xml:tm) allows revision and translation memory to be stored directly in XML documents via the XML namespace mechanism, thus making linguistic assets available to any process that requires them. A relatively new standard, xml:tm interfaces well with DITA, TMX, and other leading standards to allow for integration of multilingual concerns during the authoring and revision phases.

- **Global information metrics Management eXchange - Volume** (GMX-V) is the first part of a proposed tripartite standard for representation of metadata about localization and translation projects. GMX-V specifically addresses the need for consistent, verifiable, and cross-platform word and character counts. In extreme cases the word counts provided by various text-processing applications can vary by as much as 30%, making any estimations of work or cost uncertain and contingent upon the tool being used. While such extreme variations are not normal, smaller variations can result in substantial cost differentials if two parties agreed to translation or other related tasks based on different assumptions about volume. While GMX-V will not replace application-specific word counts that are suitable for specific tasks, it does provide a mechanism for parties to agree upon costs up front with certainty. Forthcoming portions of the GMX standard will focus on textual complexity (e.g., grammatical and lexical complexity) and pre-negotiated quality requirements (in coordination with ISO Technical Committee 37 projects).

- The **LISA QA Model** is the most widely-used localization quality metric. Quality Assurance (QA) has traditionally been difficult in the language industries because evaluations are subjective and subject to personal preference. Thus one individual might find nothing objectionable in a translation while another might consider it very bad. With support for a variety of models (including SAE J2450 and custom error profiles), the LISA QA Model provides a way for objective quality decisions to be made.

While some of these specifications are already widely used in the globalization industry, others are more recent and are just beginning to be implemented. They all, however, are designed to provide solutions to common problems faced by anyone working with text processing, particularly multilingual text processing. A standards-based workflow designed around LISA standards can offer significant technical and business benefits to implementers by reducing dependence on specific tool or services vendors and allowing choice in linguistic tools and processes to be made based on competitive differences rather than on technical lock-in. The tutorial will describe each specification, provide examples of its use, and discuss how individuals working with text—even outside of the globalization industry—can benefit from them. This tutorial will provide a solid introduction to the standards and specifications mentioned in the description. They play an important role in the development of tools for dealing with multilingual text, but also address areas that impact any individual or organization interested in processing text: segmentation of text, storing of text histories in XML, word counts, etc.

In addition, LISA is currently in the process of migrating these standards and guidelines to an ISO framework (TBX is already being worked on as ISO 30042), so their importance to governmental bodies and any organization that uses ISO standards will only increase. By becoming familiar with these specifications at an early phase, participants will be better prepared to comment on them as they enter ISO and to implement them within their organizations.

Tutorial Speakers

Alan K. Melby
Brigham Young University, Provo
Board Member, American Translators Association (ATA)

Member, LISA Open Standards for Container/content Allowing Reuse (OSCAR) standards steering committee
email: melbyak@yahoo.com
web: http://www.ttt.org

Alan K. Melby has worked in the translation/localization industry since the 1970s, starting in machine translation, switching in the 1980s to productivity tools for human translators. In the 1990s his focus shifted to work on translation-related standards. He holds a PhD in computational linguistics from Brigham Young University under the direction of William J. Strong, director of the Acoustics Research Group. He received the Eugen Wüster Prize for contributions to the field of terminology in 2007.

Arle R. Lommel
Chair, Open Standards for Container/content Allowing Reuse (OSCAR) standards committee, The Localization Industry Standards Association
LISA delegate to ISO TC 37 and TC 46
email: arle@lisa.org
web: http://www.lisa.org

Arle Lommel has worked in the localization industry with LISA since 1997. He has been active in the development of LISA standards since 2001. He holds a BA in linguistics from Brigham Young University, Provo, where he worked with Alan K. Melby, and an MA in folklore studies from Indiana University, Bloomington. He has actively published in linguistics and has headed LISA's standards initiatives since 2003.

Kara Warburton
Terminologist, IBM
Member of ISO TC37, SC3
Member of LISA Terminology Special Interest Group

Kara Warburton has been the head of IBM’s terminology management program for ten years, where she led the development of a multilingual database and its deployment across content authoring, translation, and extended applications such as search engines. She is the head of the Canadian delegation on ISO TC37, SC3, where she has contributed to terminology standards. She also spearheaded the LISA Terminology Special Interest Group, which defines best practices for terminology management in the localization industry. She holds a BA in Translation Studies, and an MA in Terminology.

= =