Unicode Localization Data Interoperability TC Overview (ULI)

What's a word? What's a sentence? Why is this business-relevant?

Christian Lieske, SAP (Walldorf, Germany)
Helena Shih Chapman, IBM (Waltham, Massachusetts, USA)

META FORUM 2013

Context and Overview

The Unicode Localization Interoperability Technical Committee (ULI-TC) was established in 2011 with the goal of helping to ensure interoperable data interchange of critical localization-related assets. ULI's work is relevant to speech/natural language processing, analytics tokenization etc. including translation memories, segmentation rules, and more. What ULI is building forms the foundation of many other downstream technologies: memory interchange, speech/natural language processing, analytics tokenization etc.

META-FORUM 2013 – Connecting Europe for New Horizons
Christian Lieske, SAP (Walldorf, Germany), Helena Shih Chapman, IBM (Waltham, Massachusetts, USA): Unicode Localization Data Interoperability (ULI) Technical Committee Overview

Unicode & Segmentation (1/3)

- More than a character repertoire – an **ecosystem**, a **stack of standards**
- Parts of the ecosystem are related to “segmentation” questions such as “How can text entities such as **sentences** be broken down into sub-entities such as **words**?”
- Segmentation is important for **business analytics** and **translation**…

Unicode & Segmentation (2/3)

Most prominent members of the Unicode ecosystem related to segmentation:

- Unicode Text Segmentation report

  TR#29 http://www.unicode.org/reports/tr29

- Unicode Line Breaking Algorithm

  TR#14 http://www.unicode.org/reports/tr14

- Common Locale Data Repository

  CLDR; see http://cldr.unicode.org

Unicode & Segmentation (3/3)

Comprehensive support for Unicode is provided by the International Components for Unicode (ICU, www.icu-project.org), a **software library** used in many applications.

ULI Credo

If Unicode and its “citizens” CLDR, and ICU get segmentation right, many applications get text processing right:

- Business analytics
- Speech/natural language processing
- Memory interchange
- Sorting
- Searching
- …

ULI Scope & Objectives

- **Gather requirements** for core and extension of the standards in the area of text segmentation and content memory
- **Establish core specification scope**, extension domain, and reference implementation to improve the usefulness of existing standards
- Create a repository of reference user profile and scenarios to **demonstrate interoperability** across desired standards
- Provide **consistent interpretation of the specification**, extension and profiles

ULI Setup

Logistics

- Meet once a month by telephone
- Regular participation by IBM, Microsoft, Yahoo, Google, SAP, Globalization and Localization Association (GALA), and XML Localization Interchange File Format Technical Committee (XLIFF TC)

Challenges

- Need more translation tool vendor involvement
- Solicit additional participation from key industry conferences

Open for participation

- Active participation is expected
- Need to be a member to attend meetings regularly
- For details, see TC Procedure on Unicode site

ULI 2012

Internal agreement on plain text content boundary joining and separate best practices:

- Leveraging TR#29
- Agreed syntax for referencing CLDR elements (XPATH to the CLDR parent element level; initially vetted
  English, German, Russian, and Spanish – see http://unicode.org/uli/trac/browser/trunk/abbrs)
- Demoed behavior of updated ULI input (see http://demo.icu-project.org/icu-bin/icusegments

ULI 2013/2014

- Draft implementation to demonstrate ULI progress
- CLDR and ICU contribution integration:
  - Initial ULI input for sentence level segmentation submitted to CLDR 24 due September 15, 2013 (see http://cldr.unicode.org/index/downloads/cldr-24)
  - Plugin implementation to ICU in progress for ICU 52 due October 2013 (see http://site.icu-project.org/download)
- Open source Computer-Assisted Translation integration in 2014 (ongoing evaluation of ICU implementation, based on ULI input into OpenTM2, see http://www.opentm2.org)