

Localisation, Centre for Next Generation Localisation and Standards

Dimitra Anastasiou

Localisation Research Centre,
Department of Computer Sciences and Information Systems,
University of Limerick, Ireland

Abstract

Localisation is the adaptation of a product to a language-culture combination, called target *locale*. Software and website localisation are the two most common types of localisation; many industries and particularly, language service providers (LSPs) offer such services.

In this article, though, we will focus on localisation in Academia and particularly on a project called "Centre for Next Generation for Localisation" which currently runs in Ireland.

The second part of the article will describe standards in terms of localisation. The standard XLIFF will be described in more detail in section 5.2.

Keywords: Centre for Next Generation for Localisation, globalisation, internationalisation, locale, localisation, Localization Industry Standards Association, Localisation Research Centre, standards, XLIFF

1. Introduction

Localisation, internationalisation and globalisation are terms which will appear throughout the article and thus we define them in Section 2. We provide the definitions given by Localization Industry Standards Association¹ (LISA) and describe them by focusing on the individual words included in the definitions.

Section 3 refers to the Localisation Research Centre (LRC) which is based at the University of Limerick in Ireland and is the educational centre for the localisation industry. In section 4 we describe one of the projects of the LRC, the Centre for Next Generation for Localisation (CNGL) and the three main challenges which are

¹ <http://www.lisa.org/>, 27/09/09

commonly faced within a localisation project and are addressed by CNGL, i.e. access, volume and personalization. Section 5 is about localisation standards and more precisely, we refer to two standards developing organizations, namely LISA and OASIS and some of their certified standards.

2. Definition of terms

In this section we provide the definitions of localisation, internationalisation and globalisation according to the Localization Industry Standards Association (LISA). LISA consists of over 200 corporate clients and their globalization solutions partners and provides business guidelines and multilingual information management standards (see section 5).

We start by the definition of localisation:

"Localisation involves taking a product and making it linguistically and culturally appropriate to the target locale where it will be used and sold."

If we analyse each word of the aforementioned definition, localisation deals with products. Product can be either a software, a tool or a market item in general. Apart from products, localisation also deals with services, such as websites, e-mails, instant messages and so on.

"Making the product linguistically appropriate" essentially means translation. However, there are other linguistic aspects to consider here. To mention only one example, the source text has to be linguistically correct in order for the translation to be performed smoothly in terms of quality, volume, cost and time, i.e. high volume in high quality with low cost in low time. This linguistically appropriate source text should be prepared by authors who are specialised in authoring and technical documentation. Controlled Language plays also a significant role here. According to Muegge (2007), a controlled language has the two following essential characteristics:

"The grammar of the controlled language is typically more restrictive than that of the general language, and the vocabulary of the controlled language typically contains only a fraction of the words that are permissible in the general language. This means that authors who write in a controlled language have fewer choices available when writing a text."

We will now discuss the second part of the definition: "making the product culturally appropriate". This means that the images, the symbols, the colours and generally the visualisation of the product shall be in accordance with the culture of the

target locale. It is not an uncommon situation that different colours, icons and even the alignment of text are changed, when one switches to a website from one available language to another.

We now describe the third and last part of the definition: "appropriate to the target locale where it will be used and sold". Target locale, as aforementioned, is the combination of language and culture. Locale is not necessarily a country, as in one country there can be more than one languages (official and spoken) and accordingly, more than one culture. In this target locale the product will not be only used, but also sold. That implies a plethora of marketing enterprises behind the localisation concept. This is exactly where globalisation comes into play.

All these parts of the definition that we just analysed are indispensable in order to accomplish a successful localisation project.

Now we focus on internationalisation which is the step before localisation. LISA provides the following definition for internationalisation:

"Internationalisation is the process of enabling a product at a technical level for localisation."

The important part of this definition is "enabling a product at a technical level". This is the engineering and designing stage of a product. Localisation engineers are responsible for correctly engineering localised content in software applications, help systems, websites and documentation. Each type of content presents its own unique challenges. For example, engineering localised software involves compiling and executing the application, resizing and the rearrangement of user interface elements, while engineering localised help systems involves compiling and executing the help system, checking the page layout, text formatting and validating navigation elements such as table of contents, index, browsing sequence and hyperlinks.

As far as globalisation is concerned, it is a process that includes both localisation and internationalisation and not only these. LISA defines it as follows:

"[Beyond localisation and internationalisation], globalisation encompasses global and local marketing, local after-market support, the establishment of local business presence, and many other aspects of global business."

We see that marketing and business are keywords in this definition. A product which has been successfully designed – task of internationalisation – and come to the target locale – task of localisation – has not necessarily become global. In order to

become global, it has firstly to be attractive and catchy to gain the attention of the locals, secondly to be sold in amounts that returns a profit to the selling company and thirdly to remain among the selling product preferences at the target market. The target locale should have their needs and wants satisfied by the localised products and services and also pay for them establishing them as a standard preferred market item; this is the task of globalisation.

3. Localisation Research Centre

In 1995 the Localisation Research Centre² (LRC) was established as the Localisation Resources Centre at University College Dublin (UCD). In 1999 it was re-constituted as the Localisation Research Centre (LRC) and moved to the University of Limerick (UL). The LRC comprises a director (Reinhard Schäler), faculty members, staff and research students.

It is regarded as the information, research and educational centre for the localisation industry, as its activities include development and evaluation of localisation tools, the establishment of a Localisation Tools Library, consultancy services, education and training. Since 2009 the LRC has offered a Master about Global Computing and Localisation³. It also annually publishes a journal called "Localisation Focus - The International Journal of Localisation"⁴ which is the only peer reviewed and indexed academic journal focusing exclusively on localisation and the localisation industry.

The LRC cooperates with worldwide digital publishers and their partners who are interested in future technologies and processes for Globalisation, Internationalisation, Localisation and Translation (GILT); it also co-operates with researchers and students, the media, consultancy firms, government agencies, and the European Commission.

A number of EU projects have been managed by the LRC, such as IGNITE, Transrouter and CLP⁵ (The Certified Localisation Professional programme). CLP has been successfully implemented by The Institute of Localisation Professionals (TILP)

² <http://www.localisation.ie/>, 23/08/09

³ <http://www.csis.ul.ie/course/LM632/>, 23/08/2009

⁴ <http://www.localisation.ie/resources/locfocus/index.htm>, 23/08/09

⁵ <http://www.localisation.ie/resources/Research/clp/index.html>, 23/08/2009

and was officially launched in 2004 at the annual LRC Localisation Conference. CNGL (Centre for Next Generation for Localisation) is the most up-to-date project and we will describe it in section 4.

4. Centre for Next Generation for Localisation

Here we present the Centre for Next Generation Localisation⁶ (CNGL), a dynamic Academia-Industry partnership project which has been in operation since 2008. There are over 100 researchers developing novel technologies which address the localisation challenges of volume, access and personalisation. We would like to address these three issues in detail in subsection 4.1.

The director of the CNGL project is Josef Van Genabith. There are four academic Irish university partners: Dublin City University (DCU), Trinity College Dublin (TCD), University College Dublin (UCD) and University of Limerick (UL). As for the industrial partners, we refer here only to some: Alchemy, DNP, IBM, Microsoft, SDL, SpeechStorm, Symantec, Translan, VistaTEC. All industrial partners listed as well as more information about the CNGL project can be found in the CNGL annual report 2008⁷. In the following paragraphs we summarize the research tracks within CNGL.

The CNGL research is based on the four following technologies:

1. Integrated Language Technology;
2. Digital Content Management;
3. System Framework and
4. Localisation

Research into the Integrated Language Technology (ILT) includes Machine Translation (MT), Speech Technologies and Text Analytics. The examination and improvement of the quality of the state-of-the-art MT systems with focus on hybrid and syntax-based MT systems, word alignment and controlled language are some of the goals of the ILT research group. MATREX (Du et al., 2009) is a multi-engine MT system, developed at DCU, which combines example-based with statistical MT aspects.

⁶ <http://www.cngl.ie/>, 23/08/2009

⁷ http://www.cngl.ie/Press/CNGL_Annual_Report_2008.pdf, 23/08/2009

As for the Speech Technologies, the major focus is to implement speech recognition and synthesis within a MuSE Speech Technology platform. Text analytics includes among other things, text classification, alternative category labels and datasets.

The Digital Content Management group focuses on the combination of Adaptive Hypermedia and Information Retrieval and Extraction. The user queries should be tailored according to the users' requirements, e.g. their goals, abilities, interests, knowledge and so on. The dynamic adaptation of ad hoc queries for Next Generation Localisation, the automated generation of metadata and the dynamic composition of intelligent responses are some of the research areas.

The System Framework group combines Integrated Language Technology and Digital Content Management. They deal with the demonstrators of the system and aim to create a framework for evaluation of core technologies. The researchers develop a workflow management system that supports the adaptive integration and operation of services and define an architecture within various components can be dynamically integrated with user applications.

As far as the localisation research track is concerned, the LRC (see section 3) which is based at the UL has divided the research into three sub-tracks: i) multilingual digital content development, ii) translation, adaptation and evaluation and iii) localisation workflows. The first one, multilingual digital content development, focuses on the design of the original content having localisation in mind as well as the creation and maintenance of the digital content right from the design stage. It has been working on a Localisation Knowledge Repository (LKR) (see Ryan & Anastasiou, 2009), where internationalisation guidelines are incorporated into the content development process. The LKR consists of a Digital Library, Test Area and Virtual Community. The Test Area facilitates automating elements of the authoring, enabling and pre-translation testing processes. The Virtual Community enables users to connect with other content developers, upload and download resource files, and post discussion forums. Also, a localisation taxonomy, which defines the file formats of localisation relevant data, is currently built.

The second research group works on translation, adaptation and evaluation by focusing on the current tools and technologies and exploring whether they cover the requirements of the localisers. The third group is about localisation workflows and

examines localisation scenarios with workflow ranging from bulk to personalised localisation.

4.1 Challenges

The three main challenges which are addressed within CNGL are volume, access and personalisation. All interrelate with each other and thus localisers have to find the balance between these challenges in order to have the desired result. We describe these essential challenges in the following paragraphs.

Access relates with the personalised multilingual social networking. Social networking is the grouping of individuals who are connected with each other through friendship, professional relationship, shared knowledge or common interests about an activity, such as sports, arts, etc. The social networking sites which gained ground the last years are these ones where the people can write articles in a free encyclopaedia (www.wikipedia.org), upload pictures and share them with their friends (www.flickr.com), upload own videos (www.youtube.com) and so on, i.e. the user ceases to be a passive recipient, but becomes an actor instead. All these are included under the umbrella of "Web 2.0". The concept of "Web 2.0" began with a conference brainstorming session between O'Reilly and MediaLive International. According to the Web 2.0 map presented at the O'Reilly Media, Inc.⁸, the Web is used as a platform and the users can control their own data. Nowadays services replace packaged software and collective intelligence is harnessed.

Another term which was first coined by Jeff Howe in June 2006 *Wired* magazine article⁹ is *Crowdsourcing*. The term has become popular with business authors who wanted to achieve business goals by leveraging the mass collaboration enabled by Web 2.0 technologies. More information can be found in Howe's (2008) book "Why the Power of the Crowd Is Driving the Future of Business".

As for the second challenge, personalisation, it deals with the personalised production content for informal learning. It takes into account the user requirements and thus the user interfaces should be designed in not only a functional, but also in a

⁸ <http://oreilly.com/web2/archive/what-is-web-20.html#mememap>, 23/08/2009

⁹ <http://www.wired.com/wired/archive/14.06/crowds.html>, 23/08/2009

more friendly and catchy way and should be adapted to the user needs and preferences.

Volume is the challenge which is mainly addressed by bulk localisation, i.e. localisation for big publishers and corporate institutions. The content nowadays is localised in ever more languages and this should be necessarily considered in relation to the volume.

The objective of CNGL is to counterbalance these challenges, access, volume and personalisation by examining the state-of-the-art and developing new localisation technologies.

5. Localisation and Standards

The existence of standards in localisation is necessary and fundamental. Tool providers shall agree on a certified standard, so that localisation industries have the flexibility to change their tools. Also, freelance translators shall not depend on a specific tool or tool provider; thus the standards should be flexible and open.

5.1 LISA/OSCAR Standards

The Localisation Industries Standards Association (LISA) hosts a special interest group called Open Standards for Container/Content Allowing Re-use (OSCAR). LISA/OSCAR develops technical standards¹⁰ for the globalization process. Some of them are the following:

- a. Term Base eXchange (TBX): it is the open, XML-based standard for exchanging structured terminological data that has been approved as an international standard by LISA and ISO. TBX, assuring controlled terminology, improves the quality of the localised text, reduces localisation cost and brings the products faster to market.
- b. Translation Memory eXchange (TMX): the vendor-neutral open XML standard for the exchange of Translation Memory (TM) data created by Computer Aided Translation (CAT) and localisation tools. The organizations are not dependent on a specific tool or tool provider any longer.

¹⁰ <http://www.lisa.org/Standards.30.0.html>, 23/08/2009

- c. Global information management Metrics eXchange (GMX): it is a family of standards of globalisation and localisation-related metrics with three main components, i) volume, ii) complexity and iii) quality (proposed). These metrics provide standard definitions to assess the quantity, textual complexity and quality requirements for globalisation tasks.

5.2 OASIS and Standard XLIFF

In this chapter we turn our attention to another Technical Committee called Organization for the Advancement of Structured Information Standards (OASIS)¹¹ Technical Committee¹². This committee is a non-profit consortium that develops, converges and adopts open standards for the global information society.

One of the open standards is XLIFF¹³ (XML Localisation Interchange File Format). Generally, XLIFF is mainly a format for exchanging localisation data. XLIFF could be used to exchange data between companies, such as a software publisher and a localisation vendor, or even between localisation tools, such as Translation Memory (TM) systems and Machine Translation (MT) systems; thus it is tool or partner-independent.

XLIFF is developed by representatives of all aspects of the translation industry and has standardised methods for automating workflow, translation word counts, TM and segmentation. It is intended to give any software provider or technical communicator a single interchange file format that can be understood by any localisation provider.

The benefit of XLIFF is that it enables translators to concentrate on the text to be translated. It provides tags and attributes for review comments (metadata) as well as the translation status of individual strings. It also separates localisable text from formatting. However, the most important advantage of XLIFF is that it allows many separate tools to work on files. While working on source file formats, it is not easy for users to run both a MT and a TM on the same files. Moreover, during review it would be difficult to distinguish which part of text came from the MT tool, the TM tool, or a

¹¹ <http://www.oasis-open.org/who/>, 23/08/2009

¹² http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=xliff, 23/08/2009

¹³ <http://docs.oasis-open.org/xliff/xliff-core/xliff-core.html>, 23/08/2009

human translator. By contrast, if the tools, MT and TM, read the XML-based standard XLIFF, they could be run sequentially.

Summarising, XLIFF enhances localisation research by coping easily with localizing different types of source files, providing a common platform for localisation tools vendors and facilitating the localisation process through its commenting features and metrics.

6 Summary

In this article we introduced and explained the definitions of localisation, internationalisation and globalisation given by LISA. Then we referred to the main activities of an educational centre based at the University of Limerick, the Localisation Research Centre (LRC). The current project, CNGL, in which the LRC together with three other Irish Universities is involved, was highlighted and the various technologies researched were described. Within CNGL, it has been examined how the three main challenges which are faced within a localisation project, i.e. access, volume and personalisation can be counterbalanced. The last part of the article briefly explained the importance of standards in localisation and referred to two standards developing organizations, i.e. LISA and OASIS and some of their certified standards.

Bibliography

Du, J., He, Y., Penkale, S. & Way, A. (2009), 'MATREX: The DCU MT System for WMT 2009'. In: *Proceedings of the 4th Workshop on Statistical Machine Translation*, EACL 2009, Athens, Greece, 95-99.

Howe, J. (2008). *Crowdsourcing, Why the Power of the Crowd Is Driving the Future of Business*. Crown Business.

Muegge, U. (2007), "Controlled language: the next big thing in translation?". In: *ClientSide News Magazine* (ClientSide Publications) 7 (7): 21–24. <http://www.translationdirectory.com/articles/article1359.php>.

Ryan, D. & Anastasiou, D., (2009), 'Developing Digital Content for Global Audiences, in: *tcworld Conference Proceedings*, Wiesbaden, Germany.