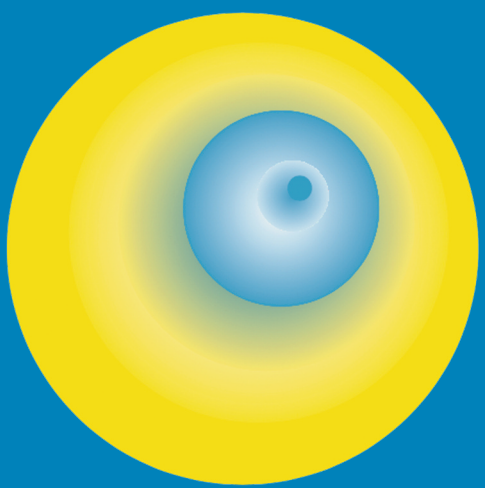
# Localisation Focus

THE INTERNATIONAL JOURNAL OF LOCALISATION

ISSN 1649-2358



Issue Sponsored By

IGNITE

The peer-reviewed and indexed localisation journal

VOL. 5

# EDITORIAL BOARD



# PUBLISHER INFORMATION

**Editor: Reinhard Schäler**, *Director*, Localisation Research Centre, University of Limerick, Limerick, Ireland
**Production Editor: Karl Kelly,** *Manager* Localisation Research Centre, University of Limerick, Limerick, Ireland
**Published by: Localisation Research Centre**, CSIS Department, University of Limerick, Limerick, Ireland

# AIMS AND SCOPE

**Localisation Focus – The International Journal of Localisation** provides a forum for localisation professionals and researchers to discuss and present their localisation-related work, covering all aspects of this multi-disciplinary field, including software engineering, tools and technology development, cultural aspects, translation studies, project management, workflow and process automation, education and training, and details of new developments in the localisation industry. Proposed contributions are peer-reviewed thereby ensuring a high standard of published material. Localisation Focus is distributed worldwide to libraries and localisation professionals, including engineers, managers, trainers, linguists, researchers and students. Indexed on a number of databases, this journal affords contributors increased recognition for their work. Localisation-related papers, articles, reviews, perspectives, insights and correspondence are all welcome.

To access previous issues online go to http://www.localisation.ie/resources/locfocus/pdf.htm and click on the issue you wish to download. Use the following logon details - username: locfocsub and password: V610808

Members of **The Institute of Localisation Professionals (TILP)** receive Localisation Focus – The International Journal of Localisation as part of their membership benefits. Membership applications can be filed electronically from ***www.tilponline.org*** Change of address details should be sent to LRC@ul.ie

**Subscription:** To subscribe to Localisation Focus - The International Journal of Localisation visit www.localisationshop.com (subscriptions tab). For more information visit www.localisation.ie/lf



**Localisation Focus – The International Journal of Localisation** (ISSN 1649-2358) is published and distributed annually and has been published since 1996 by the Localisation Research Centre, University of Limerick, Limerick, Ireland. Articles are peer reviewed and indexed by major scientific research services.

# FROM THE EDITOR

## I

Egypt's Minister of Communications and IT recently announced an agreement reached between the LRC, Egypt's Ain Shams University (with 174,000 students) and his own Department to jointly develop localisation courses in Egypt. The minister supports effort by the industry to develop Egypt as the localisation hub for the Arabic world and North Africa. Following the agreement reached earlier with the University of Florianopolis (Brazil) and ongoing negotiations with universities and state-sponsored bodies in Asia and Africa, this development highlights the enormous interest in development localisation and the efforts undertaken by the signatories of the Limerick Declaration, establishing the Global Initiative for Local Computing (GILC).

Much has been reported about the merger of the giants, Lionbridge and Bowne Global Solutions. Much more interesting though than the merger is the medium- to long-term effect this merger will have on the rest of the industry. In this issue, we report on another merger, on a smaller scale, but no less significant. Welocalize and Connect Global Solutions have decided to combine their forces to fill a void left following the giant leap taken by Lionbridge and Bowne. In our interview with E. Smith Yewell of Welocalize, Smith points towards an important factor, often forgotten by those with an almost exclusive focus on shareholder value: This merger, he says, was "a merger tailored to meet a market opportunity but also one tailored to staff and client needs."

The LRC has just finalised a detailed review of its activities and its position in the localisation landscape. This review was supported by its Industrial Advisory Board, industry leaders, state bodies and distinguished academics, and facilitated by Dublin-based company Product Innovator Ltd. As a result of this review, the LRC has intensified its relationship with worldwide digital publishers and their partners who are interested in future technologies and processes for GILT. The LRC will continue to provide relevant well-researched content-rich information on future trends and technologies, offering a unique platform for industry and academic collaboration and providing an unparalleled network of expertise. One important tool for this work is Localisation Focus – The International Journal for Localisation, which has just been included in two highly prestigious academic indexes, Ulrichs Periodicals Directory (see http://www.ulrichsweb.com/ulrichsweb/) and INSPEC (see http://www.engineeringvillage2.org), highlighting the relevance and quality of our publication.

***Reinhard Schäler, March 2006***

## II

Welcome to the newly designed Localisation Focus – The International Journal of Localisation!

With the new title page and the improved layout, we wanted to add visual impact to our efforts to provide the localisation community with a high-quality international journal publishing peer-reviewed articles that are now indexed by major international scientific services — thanks to the combined efforts of our Editorial Board, editors, designers, and peer-reviewers.

The next steps in the development of our journal will bring you a greater coverage of activities of the Global Initiative for Local Computing (GILC), a dedicated review section (books, events, courses), and an improved regional distribution network for Asia, Africa and South America.

This year marks the 10th Anniversary of Localisation Focus. To celebrate this occasion, Thomas Keogan of the LRC has prepared a special, limited edition of all issues published over that period which can be ordered directly from the LRC and online at www.localisationshop.com.

The Localisation Research Centre (LRC) is continuing to work with partner organisations globally to develop and facilitate local computing. In June 2003, the LRC established the Localisation Tools and Technology Laboratory and Showcase (LOTS) to facilitate easy access to state-of-the-art localisation tools and technologies for practitioners, researchers, teachers and students.

This effort was funded by the European Union eContent Programme and supported by the localisation tools industry. On 19 April 2006, the LRC launched the GILC-LOTS Satellite Distribution at

the International Symposium on ICT for Rural Development in Kuching, Malaysia, and I handed the first copy of this distribution to Prof. Dr. Abdul Rashid Abdullah, the vice chancellor of the Univerisiti Malaysia Sarawak, following the signing of a Memorandum of Understanding between the LRC and that university.

This was made possible, to no small degree, thanks to the efforts of Dr. Alvin Yeo of the University of Malaysia.

I also would like to thank Michael Bourke of the LRC, as well as Alchemy, CatsCradle, PASS, Project Open, Aquino and SDL/TRADOS who have contributed their tools free-of-charge to this distribution. The GILC-LOTS Satellite Distribution, with each installation worth tens of thousands of euros, will be rolled out by the LRC free-of-charge initially to ten partner universities worldwide.

***Reinhard Schäler, June 2006***

## III

The 11th Annual LRC Conference, IGNITE: The Localisation Factory (25–26 October), will take place at the European Foundation in Dublin. Two highly distinguished keynote speakers will address delegates: Lisa Moore (IBM's Instant Messaging (IM) Products Globalisation Manager, USA) and Richard Ishida (Internationalisation Activity Lead, World Wide Web Consortium – W3C, UK). They will be joined by a large number of international speakers, covering topics such as translation quality and automation and process automation.

One important focus of the conference will be the demonstration of initial findings of the European Union–funded IGNITE project. While some proprietary automated localisation environments are already operational within some large multinationals, IGNITE will open large-scale process automation up to the industry as a whole.

A special Ask the Experts session, organised by The Institute of Localisation Professionals (TILP), will take place at this year's 28th ASLIB Conference, Translating and the Computer, in London's Copthorne Tara Hotel, Kensington (16–17 November). At this session, entitled TechLink: Education and Training for Localisation, international experts representing the freelance, client, vendor and educational sectors will share their verdict on the current situation and propose future strategies. Attend this TILP event and voice your opinion, contribute with your expertise and influence current and future training and educational programmes in our industry.

In July of this year, the LRC signed a Memorandum of Understanding with the University of South Africa (Unisa), the country's largest university and one of the largest distance learning universities in the world. Dr Wallmach of Unisa said

*"South Africa alone has 11 official languages, with English ranking fifth as a mother tongue. Only 22% of South Africans fully understand English and there is not only a legal obligation, but also a social responsibility to make digital content available to South Africans in their own language. We are very excited about our plans to collaborate with the LRC at the University of Limerick and are already about to set up a satellite version of their Localisation Tools and Technology Laboratory (LOTS) for our students."*

Following the agreements with GeNESS, the Brazilian-based Research Centre at the Universidade Federal de Santa Catarina (UFSC) and the Universiti Malaysia Sarawak (UNIMAS) at Kuching in Malaysia, this is the third Memorandum of Understanding signed by the LRC with leading universities active in localisation-related research. Negotiations are currently under way with institutions in India and the USA to broaden this unique network of academic research and educational organisations.

***Reinhard Schäler, September 2006***

**Production Editor's Note**

This is a special, distilled, edition of Localisation Focus The International Journal of Localisation that collates the peer-reviewed papers that were published in Volume 5 Issues 2 and 3, the final peer reviewed academic content that was published, before the format change in 2007. This edition updates these papers into the format used from 2007 onwards.

The original issues are, of course, still available to download and read from www.localisation.ie and they contain extra content in the form of industry news and project updates.

***Karl Kelly, June 2017***

# Lessons Learnt in the Development of Applications for Remote Communities

**Alvin W. Yeo, Azman Bujang Masli, Siou-Chin Ong,**
**Peter Songan, Jayapragas Gnaniah, Khairuddin Ab Hamid, Poline Bala**
**Universiti Malaysia Sarawak (UNIMAS)**
**94300 Kota Samaharan**
**Sarawak**
**Malaysia**
**alvin@fit.unimas.my**

**Abstract**

In this paper, we highlight lessons learnt from our experience in the development of three applications for two small remote communities in Sarawak, a state of Malaysia, which has been provided with access to information and communication technologies. The applications developed include a digital library employed to capture oral traditions of the Kelabits; a website to promote tourism in Bario, and a word processor localised to accommodate the Kayan language. The Kelabits and Kayans are two of 27 ethnic groups found in Sarawak. These lessons highlighted will be discussed vis-à-vis the technological, operational, logistical, and strategic aspects of systems development for remote communities.

**Keywords:** *Bridging the digital divide, eBario Project, ICT, Open Source, digital library, word processor, tourism website, community informatics*

## 1. Introduction

There are many projects aimed at bridging the digital divide that have been deployed all over the world — specifically in rural areas. Such projects have been implemented in the hope of bringing about the many potential benefits to these communities, in particular, to improve their social, economic and cultural well-being. Instances of these benefits include the elimination of the barriers to physical and virtual isolation, providing access to available information, and increasing opportunities to expand businesses to reach new markets. While many of these projects report on the provision of access of Information and Communication Technologies (ICTs) to remote communities, there are few studies that report specifically on development and its processes of specific technologies for the rural communities.

Thus, this paper aims to provide a description of and details on the lessons learnt from the development and processes employed in the implementation of three applications or systems. In the next section, this paper will provide details of the location of the two remote communities i.e., Bario and Long Bedian — the communities for which the applications were developed. In addition, the eBario project, which aims to bridge the digital divide, is also described. This project provides the context within which the three applications were developed. The ensuing section will then described the applications that were built — in particular, the rationale for the development, the methodology employed, and the outcomes of the implementation. The lessons learnt are detailed in the form of an examination from technological, operational, logistical, and strategic aspects.

## 2. Bario

Bario is located in the Kelabit Highlands, near the Kalimantan and Sarawak border (see Figure 1). It is the 'unofficial capital' of the 'land' of the Kelabits, one of the 27 ethnic groups in Sarawak. Prior to the introduction of a daily flight into Bario, the Kelabits only means of communication with the closest town was by foot — climbing mountains, following mountain ridges, and crossing and re-crossing rivers and valleys for several weeks. Today, flying (which takes about an hour) to Bario, the main Kelabit centre, is the only practical way to get there.

Bario has a number of government offices, and also provides education and health services to the Bario community and surrounding villages.

There are about 1,200 people living in Bario. The Bario district is occupied principally by the Kelabit (78%) (one of the smallest ethnic groups in Sarawak), with other ethnic groups including Penan, Kenyah, Iban, Bidayuh and Malays, Chinese, as well as some Indonesian immigrants. The majority are farmers (93%), planting wet rice as their main crop. About 5% of the population work in government offices, whereas about 2% operate personal business-

es and trading. In addition to rice cultivation, the community also rears livestock such as buffalo, cattle, sheep, chicken and pigs. Some members of the community are also involved in hunting, fishing and forest gathering.

## 3. eBario: Bridging the Digital Divide

The idea of bringing the Internet to Bario was conceived as a research project to determine opportunities for social development available from the deployment of information and communication technologies (ICT) within remote communities in Sarawak. Desirable results from pilot studies in other developing countries have encouraged the team to work among those communities in Sarawak to have equal access to ICTs, specifically, the Internet which could provide significant improvements in their lives. This was included in the eBario project. Basically the goals of the eBario Project were to:

- Define the extent that contemporary ICTs can deliver sustainable human development and significant improvement to the lives of the community
- Demonstrate how significant and sustainable development can be achieved by remote communities through the innovative use of ICT

The objectives of the eBario project included to:

- Empower the Bario community to be able to employ ICTs to improve their livelihood through a people-centred/participatory approach
- Provide the Bario community and school children with access to ICTs through:
  - a computer laboratory at SMK Bario
  - a community telecentre at Bario

As part of the eBario project, numerous areas were identified as potential beneficiaries from the introduction of ICTs. These areas included education, culture, commerce, agriculture, health, community, technology, and human resource development.

### 3.1 Why Bario?

While there were many communities in Sarawak that satisfied the criteria for choosing a rural remote location, Bario was selected because of its isolation. In addition, it has basic infrastructure (no 24-hour electricity supply, gravity fed water) and no telecommunication service. This can be considered a real case of 'digital divide' and 'digital poverty'. Lastly, the community's readiness to participate, given that Universiti Malaysia Sarawak (Unimas) has had conducted other research projects in the area and thus are known to the local Bario community. Because of its remoteness, the catch-phrase was that if you could successfully implement such a project in Bario, you could do so anywhere.

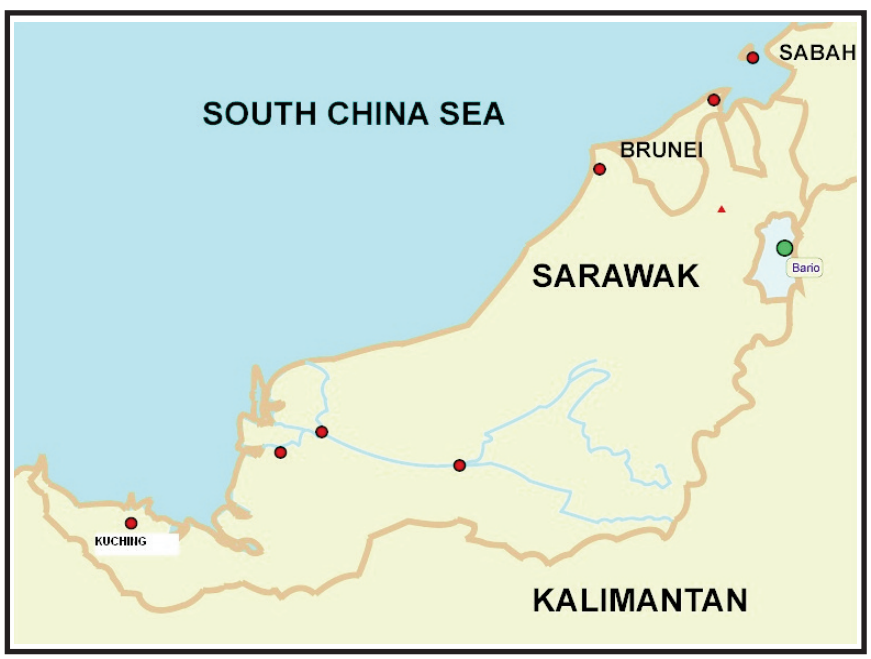



**Figure 1:** Map of Sarawak showing the location of Bario

### 3.2 Benefits to Bario Community

Numerous benefits were realised in the areas of education, and commerce. With the community's access to ICTs, there is increased computer literacy among the students, teachers and members of the community. Students from Bario are no longer disadvantaged when they go to the urban areas to continue their studies — they would be just as adept at using computers as their urban counterparts.

The community is able to communicate with the rest of the world due to the availability of telephones and Internet (via VSATs). The community, especially those involved in tourism, have taken advantage of ICTs — they are able to communicate with potential tourists directly via email, and confirm accommodation bookings online.

At the state level, the project has served to sensitise the State Government towards the potential for ICT-induced rural development. In particular, it has demonstrated the importance of ICTs to isolated communities that are denied other forms of infrastructure. The Government of Malaysia is paying increasing attention to rural development; different rural ICT programmes have been run — such as the *Pusat InfoDesa*, and *Medan InfoDesa*.

## 4. Long Bedian

Another remote community involved are the Kayans who live in another isolated remote location known as Long Bedian. Long Bedian is located in the Apoh Tutoh region of the Baram district, in the Miri Division of Sarawak (see Figure 1). The village comprises 180 houses and has a total population of 1,686 people. There are only two ways to get to Long Bedian from Miri town — either an express boat journey of seven hours followed by an hour-long drive to Long Bedian, or a four-and-a-half hour drive (by 4WD) from Miri. The village functions as a trading centre for the nearby villages, particularly for the Penan community. It also provides education and health services to the Long Bedian and Penan community.

The Long Bedian community comprises several ethnic groups — such as Kayan, Kelabit, Kenyah, Morek, and Punan. The Kayans are the biggest group in the village making up 95% of the population, while the Kelabits make up 3.9%. The remaining 1.1% of the total population in Long Bedian comprises the Kenyahs, Moreks and Punans. The Long Bedian community are all Christians. The population of Long Bedian consists of about 1,686 people.

The primary occupation in the Long Bedian community is farming (68.4%) planting paddy, oil palm trees, pepper, and other crops. About 5.2% of the people are government servants, with the remainder involved in either small businesses or the private sector.

The next section details the systems developed for the communities in Bario and Bedian.

## 5. Software Applications Developed

There are three applications that will be covered in this paper, namely, the *Bario Lakuh Digital Library*, a *Tourism Website* and a *Word Processor*. These applications were the outcome of three sub-projects funded by the Universiti Malaysia Sarawak.

### 5.1 Bario *Lakuh* Digital Library

This digital library project was aimed at preserving a

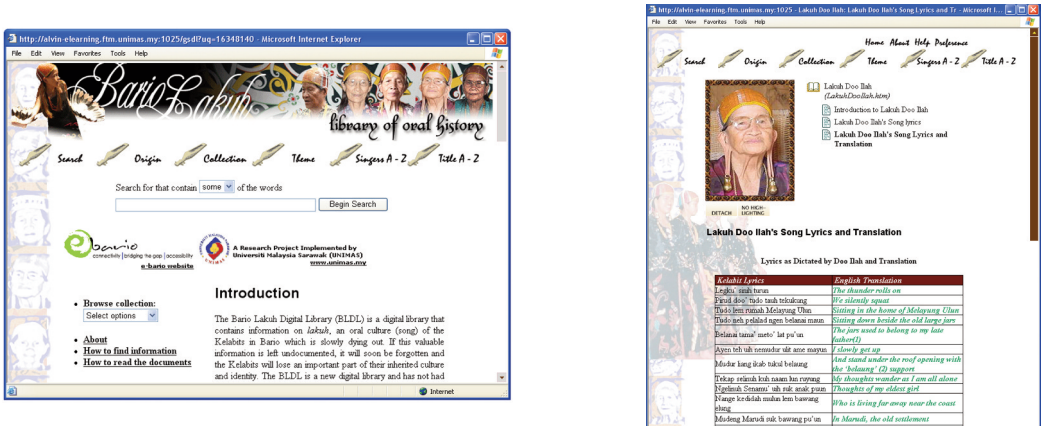



**Figure 2:** Screen shots of Bario Lakuh Digital Library

Kelabit oral tradition i.e., the traditional Kelabit songs known as '*Lakuh'*. The *lakuh* is a means of passing information about significant events on to the next generation, as well as depicting one's feelings.

Thus, one of the objectives of the Bario *Lakuh* Digital Library (BLDL) project was to explore the cultural benefits of ICT in stimulating the production, protection and popularisation of Sarawak rural communities' oral traditions, which constitute part of an indigenous knowledge system. This project, in line with the objectives of the eBario project, aimed to record and transcribe some of these traditional songs, particularly the *lakuh* songs.

There were three main phases to the project; Data collection, *Lakuh* Translation and Documentation, and Building the Digital Library.

**Data Collection and Translation:** Both audio and video recordings of the *lakuh* singers were carried out by the researchers. As the Kelabit women were only fluent in Kelabit (and spoke little Bahasa Melayu or English), it was essential to have a Kelabit speaker present. During these recording sessions, the singers were also interviewed. After recording the *lakuh*, it was transcribed and translated into English by Florence Apu, a qualified translator who is fluent in both written and spoken Kelabit as well as English. This translation was conducted in Bario.

**Building the Digital Library:** The next step was to digitise the audio and video recordings and store them in a digital library using the open-source Greenstone Digital Library Software (from the University of Waikato, New Zealand). This software allows more *lakuh* to be added into the existing library, if required.

**Outcome:** A prototype of the Bario *Lakuh* Digital Library was completed and has been published on the CDROM. It contains nine *lakuh* sung by five Kelabit *lakuh* singers. The *lakuh* lyrics are available in Kelabit (with a translated version in English), as well as in audio and a video recording (of the singer rendering the song). The background of the singers and details about the *lakuh* and its meaning are also provided.

Through this Digital Library, the cultural heritage can thus be preserved and the knowledge of the indigenous group can be passed down to the next generation. Linguists will be interested in the language used in the *lakuh* which has evolved over time; the lakuh are sung by women of the older generation, and thus uses (untainted) Kelabit.

### 5.2Tourism Website

The second application developed — a map-based tourism website — was developed as part of eBario to promote Bario as a tourist destination. It is believed that residents of Bario will benefit from eTourism.

The objectives in developing the map-based website were to:

- provide comprehensive information of Bario including maps
- provide information about lodges and homestays
- provide information about tourist guides and enable tourists to reserve a tourist guide in advance

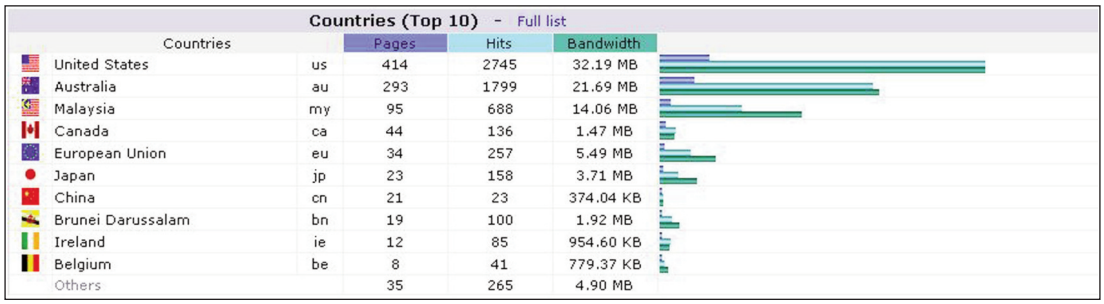
This website was developed using the web-based system development life cycle; covering web page design, framework and content development. This website also included zoom-able and interactive maps in Scalable Vector Graphic (SVG) format.

The website is complete (see Figure 3 and http://www.ebario.com). SVG was employed to provide maps of Sarawak, Kelabit Highlands, Bario Town, Pa Lungan and Pa Umur (villages in Bario). Key landmarks such as lodges, tourist attractions and government offices, such as the police station and immigration office, were also included.

Based on anecdotal evidence from visitors from Australia (on their way to Bario), the website provides the necessary information for visitors. Also, through the website, homestay owners in Bario have received emails from potential visitors enquiring about Bario. Presently, no data has been collected to determine the economic impact of the website. However, logged visits to the site show that there have been consistent numbers of visitors to the website, and not only Malaysian visitors (see 4). In eBario, the homestay owners are fully utilising the ICTs. They are using emails to contact their clients and are keen to use the Internet to promote their homestays and Bario itself.

### 5.3 Word Processor

In this project, the word processor which allowed interaction in English was customised to accommodate interactions in Kayan and Kelabit. This was implemented as part of a thesis to determine the efficacy of the existing software development lifecycle



Countries (Top 10) - Full list

| Countries | | Pages | Hits | Bandwidth |
|---|---|---|---|---|
| United States | us | 414 | 2745 | 32.19 MB |
| Australia | au | 293 | 1799 | 21.69 MB |
| Malaysia | my | 95 | 688 | 14.06 MB |
| Canada | ca | 44 | 136 | 1.47 MB |
| European Union | eu | 34 | 257 | 5.49 MB |
| Japan | jp | 23 | 158 | 3.71 MB |
| China | cn | 21 | 23 | 374.04 KB |
| Brunei Darussalam | bn | 19 | 100 | 1.92 MB |
| Ireland | ie | 12 | 85 | 954.60 KB |
| Belgium | be | 8 | 41 | 779.37 KB |
| Others | | 35 | 265 | 4.90 MB |

**Figure 3:** Top 10 countries that visited www.ebario.com (February 2016)

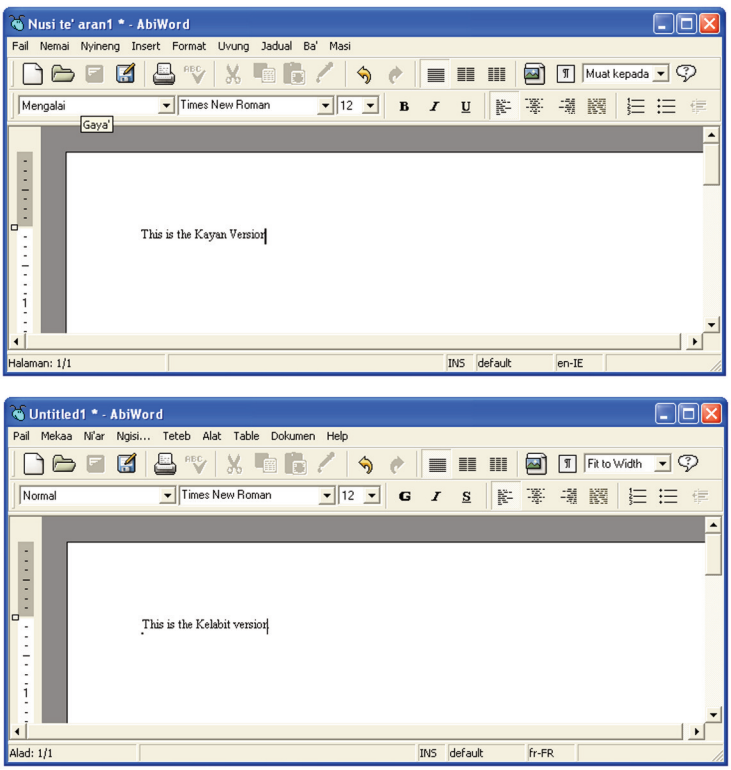



**Figure 4:** Screen Shots of the Word Processor in both Kelabit and Kayan

(SDLC); current SDLC is a Western construct and it was argued that the SDLC may need to be adapted to suit local contexts (Azman and Yeo, 2004).

Our first plan was to use the Open Source Software, *OpenOffice*. However, obtaining a build environment of OpenOffice in Windows became a major obstacle. Due to time constraints, we decided to work with a less complex software application i.e., *Abiword* which is an open source word processor.

The development was conducted in four stages. In Stage 1, we achieved a build environment to create the software, which could accommodate different languages. In Stage 2, we identified the computing terms to be translated and to translate these terms into Kayan and Kelabit (localisation phase). Translators were identified to conduct the translations; approximately 3,000 terms had to be translated in total. In Stage 3, we tested whether the Kayan and Kelabit language could be added to *Abiword*. In doing this we focused mainly on the menus and tooltips, and it was successfully carried out. The Kayan version was more complete and was evaluated by native Kayan speakers.

The results indicate that the usage by the Kayans was

similar to that experienced by first time users of software in their own language. Also those who had previously used English word processors were able to identify the English equivalent first before looking for the Kayan word. Consequently, it was difficult to measure the functionality — as the users had to translate the Kayan commands back to English.

A word processor which can accommodate Kelabit and Kayan was achieved (see 5). However, the effort in adapting OSS in the project was underestimated. The team was not aware of difficulties and only decided to adopt the less complex word processor in the middle of the project. As OSS developers are located word-wide, Internet communication was the only way to get feedback. This involved participation in mailing lists and OSS community discussions. Delays occurred as these developers are mostly volunteers, which curtails their availability to answer queries. At present standard computing terminology of Kayan and Kelabit do not exist. Thus, provision of a tool in the target language may be a way the community can preserve the language.

## 6. Lessons Learnt from the Application Development

An overview of the three applications is provided in Table 1. The overview is organised according to generic software development phases and details key activities conducted as well as activities that relate to involvement of the local community. The lessons learnt from our involvement in the development of the applications here are by no means exhaustive, but do provide guidelines for those interested in developing applications for remote communities.

### 6.1 Crucial to Form Rapport with Target Community

Forming a rapport with the target community is of immense benefit to both parties; the developers will have access to information otherwise not available elsewhere, and the local community contributes to the successful completion of the application. (Referring to **Error! Reference source not found.**, there are numerous areas whereby the locals were involved). In the case of Bario *Lakuh* Digital Library, during the data collection phase Florence Apu — a Kelabit and former English teacher — was able to identify *with whom, when, where and how* each interviews could be conducted.

### 6.2 Identify a Local Champion

Where possible, the project team members should identify a local champion who not only provides the necessary information to the project team, but also to those on-site. The local champion would act as a motivator, at the grass-roots level, to get things done. In the case of the Tourism Website, John Tarawe, was able to persuade the related parties to cooperate and provide the necessary information for the website.

### 6.3 Do Not Underestimate Logistical Problems

Travel to remote areas may impact on the scope as well as the project schedule and budget. Remoteness, long travel time and infrequent flights to such areas will increase the project duration. Also, such trips may be affected by inclement weather. For example, a flight delay in Bario due to bad weather could leave you stranded in Miri until the weather clears (which could take days).

### 6.4 Project Duration May Take a While

Another factor that may result in the development time of projects being extended is when the projects rely on volunteers. Volunteers work on a free-time basis so if they are busy with their non-voluntary work this may adversely affect task completion deadlines.

### 6.5 System Development both On- and Off-site

Given the difficulties of travel to remote areas (on-site), certain parts of the system development could be conducted off-site in order to reduce costs. Off-site implementation may be better since access to information/tools is easier than in the remote area. Similarly, usability tests can be conducted off-site if target users are available there.

### 6.6 Start Small

Where possible, applications to be developed/translated should be of a small, manageable size. Knowledge on the development environment of the target application is also crucial. Success would act as an impetus for bigger applications/projects.

### 6.7 Sustainability of Software Use

Training of the target community with the software is necessary to ensure maintenance and use of software. In the case of the word processor, besides training to use the tool, members of the community had to be trained to make minor modifications to the translations (in addition to being able to create the build environment for more major changes).

| | **Digital Library (DL)** | **Tourism Website** | **Kayan, Kelabit**<br>**Word Processor** |
|---|---|---|---|
| **Goal** | Capture indigenous oral traditions | Promote tourism | Localise a word processor |
| **Requirements Analysis** | Identify goals of DL<br>Identify the singers *<br>Identifying translator *<br>Tasks to be completed | Identify goals and requirements of websites<br><br>Identify people to interview and collect data from * | Language requirements/idiosyncrasies *<br><br>Identification of applications available * |
| **Design** | Design interfaces, navigation structure, functionalities * | Website design: interface, navigation, databases, functionalities | Design: language-dependent components are as per screen interface |
| **Implementation: Data collection** | Recording audio, video (on-site) *<br><br>Interviews (on-site) *<br><br>Transcription & translation of songs (on- and off-site) * | Recording and collection of information/content: interviews, information about culture & tourist attractions, accommodation, photos, maps (on-site) * | Translation of the language-dependent components into Kayan-Kelabit (translators were off-site) * |
| **System Implementation** | Building the libraries<br>Digitisation of the songs<br><br>Integrate into the DL<br>CD ROM produced (after evaluation) | Development of website: web pages, databases, SVG maps<br><br>Incorporation of the different media and write-ups | Modification of code to accommodate target languages |
| **Application(s) employed** | University of Waikato's Greenstone | ASP, MS SQL Server, JavaScript | Abiword * |
| **Testing and Evaluation** | Testing of the system with users [*]<br>Refinement: Editing of the translation * | System testing & usability testing [*]<br>Accuracy of information * | Usability evaluation of the word processor: corrections of the translations * |
| **Maintenance and Operations**<br><br>Current<br>Future | CD ROM distributed *<br><br>More *lakuh* to be added by local community after training provided * | Unimas currently maintains the website<br><br>Community makes changes and uploads information themselves * | Unimas makes the modifications<br><br>Community updates the information/translation * |

**Table 1:** Overview of the Three Applications Developed
On-site**: in Bario or Long Bedian;** off-site **(in Unimas); * locals involved; [*] Optional involvement of locals**

## 7. Summary

In summary, the development of software for remote communities is not just about technology or logistical issues. It is about working with and for the people. As long as the needs of the people are taken into account, the technologies (regardless of what they are) will largely be accepted — albeit with some modifications to suit the local context.

## Acknowledgements

*The authors acknowledge the funding provided by Universiti Malaysia Sarawak for the following Fundamental Research Grant projects 03/14/365/2002 (102), 01/32/381/2003 (118), and 02(01)/277/2002(15). The project team acknowledges the use of the Greenstone Software (http://www.greenstone.org) and AbiWord software (http://www.abisource.com/) and would like to thank all those involved — in particular, Bario and Long Bedian community.*

## Bibliography

Azman B.M. and Yeo, A. W. (2004). Unpublished report of Fundamental Research Grant 02(01)/277/2002(15), Building Software for Malaysians: An Application of Global-Software Development Life-Cycle.

Bala, P., Harris, R.W., & Songan, P. (2003). E Bario project: In search of a methodology to provide access to information communication technologies for rural communities in Malaysia. In S. Marshall, W. Taylor, & X. Yu (eds.), Using community informatics to transform regions (pp. 115-131). Hershey, PA: Idea Group Publishing.

Centre for International Development. Readiness for the Networked World. A Guide for Developing Countries. Retrieved July 1, 2005 from http://www.readinessguide.org.

Kano, N. (1995). Developing international software for Windows 95 and Windows NT. Microsoft Press. Redmond, WA.

Gnaniah, J., Yeo, A., Songan, P., Hushairi Z. and Khairuddin A. H. (2004). A Comparison on the Implementation Approaches for the e-Bario and e-Bedian Project. Proceedings of the Seventh Conference on Work With Computing Systems, Kuala Lumpur, Malaysia, 29 June-2 July.

Gnaniah, J., Songan, Yeo, A., P., Hushairi Z. and Khairuddin A. H. (2004). Communication Patterns of the Long Bedian Community: Implications for the Development of a telecentre. Proceedings of the Seventh Conference on Work With Computing Systems, Kuala Lumpur, Malaysia, 29 June-2 July.

Liew, J., Yeo, A., Khairuddin A. H., and Al-Khalid O. (2004). Implementation of Wireless Networks in Rural Areas. Proceedings of the Seventh Conference on Work With Computing Systems, Kuala Lumpur, Malaysia, 29 June-2 July.

Songan, P., Khairuddin A. H., Yeo, A., Gnaniah, J., and Hushairi Z. (2004). Community Informatics: Challenges in Bridging the Digital Divide. Proceedings of the Seventh Conference on Work With Computing Systems, Kuala Lumpur, Malaysia, 29 June-2 July.

Unimas. (2004). eBario Web Site. Retrieved July 1, 2005 from http://www.unimas.my/ebario.

# The Sinhala Collation Sequence and its Representation in Unicode

**Weerasinghe A.R., Herath D.L., Gamage K.,**
**Language Technology Research Lab,**
**University of Colombo School of Computing,**
**Colombo, Sri Lanka.**
arw@ucsc.cmb.ac.lk, dherath@webmail.cmb.ac.lk, kgamage@webmail.cmb.ac.lk

**Abstract**

The alphabet of a language is perhaps the first thing we learn as users. The alphabet of our mother tongue would be the first alphabet we ever learn. And yet, a closer look reveals that there is much about such an alphabet that we have not explicitly specified anywhere. The Sinhala alphabet order is a prime example. We use it, recite it and yet would be hard pressed to define it explicitly.

Sinhala is spoken in all parts of Sri Lanka except some districts in the north, east and centre by approximately 20 million. It is spoken by an additional 30,000 (1993) people in Canada, Maldives, Singapore, Thailand and United Arab Emirates. Sinhala is classified as an Indo-European language and used as an official language.

The UNICODE Collation Algorithm (UCA) is an attempt to make explicit the collation sequence of any language expressed in the UNICODE (or any other) coding system. In order to express the Sinhala collation sequence (alphabetical order) using UCA, the authors undertook the task of identifying unresolved issues facing the unambiguous definition of the order. This paper first describes the issues identified through this study, suggesting alternate solutions and recommending one of them. Finally, it sets out the recommended collation sequence for Sinhala in the form of the UNICODE collation specification. The outcome of this process is a unique and unambiguous expression of the Sinhala collation sequence which could be tested using existing tools and software environments.

**Keywords:** *UNICODE, Sinhala, Collation, UNICODE Collation Algorithm, Localisation, Internationalisation.*

## 1. Background

The Collation order of any language is one of the most important issues that has to be resolved urgently in the process of standardizing such a language. Since languages have been used for centuries by humans without worrying about their irregularities, often their constructs are extra logical. The collation sequence of most languages faces this non-logical nature. Steps are being taken to avoid these ambiguities and irregularities and also to formalize the collation sequence as much as possible since it plays a major role in the process of standardizing the languages. This is particularly so in the case of electronic texts of a language since the computer needs explicit ordering information in order to process such a language.

To illustrate simply, even in the apparently well understood case of the English Latin-1 character set, the font itself does not encode order. If it did, words beginning with 'Z' will precede those beginning with 'a' since all upper-case letters precede all lower-case letters in all encodings (including UNICODE) of English. The case for Sinhala is no exception. In fact, as will be clear in the ensuing discussion, the Sinhala collation sequence demands us to take some decisions thus far not explicitly made for the language as a whole.

## 2. Introduction to Sinhala Alphabet

The Sinhala alphabet consists of characters which represent almost all the sounds that can occur in the language. On the other hand, it is phonetically *over specified* in that there are multiple characters to represent the same sound: for example ල (dental la) & ළ (Alveolar la), න (dental na) & ණ (Alveolar na), ට (voiceless ta) & ඨ (voiced ta).
The commonly accepted *Mixed Sinhala Alphabet* has a set of sixty (60) characters. This set of characters can be classified into three categories, namely vowels, semi-consonants and consonants.

**Vowels**: There are 18 vowels in the Sinhala alphabet, අ, ආ, ඇ, ඈ,….ඕ, ඖ.
**Semi-consonants**: there 2 characters which can occur only with a vowel: ං and ඃ
**Consonants**: there are 20 consonants in the alphabet: ක, ඛ, ග,….., හ, ළ

In addition to the above characters there is another set of symbols called vowel-strokes or *'pilli',* to represent vowels sound when vowels are combined with consonants. For example: **ක** + **ා** -> **කා** (k + aa) and **ක** + **ෙ** **ා** -> **කො** (ka +o)

There is generally no disagreement regarding the order of characters within vowels, semi-consonants and consonants, except for 'ඥ' and 'ඔ'. The relative order of these character *sets* is also well defined, i.e. vowels are followed by semi-consonants which themselves are followed by consonants.

## 3. Methodology

The methodology adopted in this study is to first gather existing views and traditions of Sinhala language collation from scholars, observe the collation sequence adopted by the major standard dictionaries and to research how collation sequence is determined at various state organizations in performing their regular tasks.

It is expected that this kind of study would reveal the main issues related to the collation order of Sinhala and how these issues are addressed by scholars, in dictionaries and by organizational practice.

## 4. Issues identified in Collation Order

The following issues* were identified at the beginning of the study. These issues in mind, prominent dictionaries were searched, the views and opinions of scholars obtained, and the procedures followed by state institutions and organizations observed. The following listing identifies five major issues (first five with associated levels of importance) and three less critical ones which need to be resolved in order to proceed with the specification of an unambiguous collation sequence for Sinhala.

**Issue # 1 [Level 2]**
The positions of *'anusvara'* and *'visarga'* in the Sinhala collation sequence. While this is not really ambiguous as far as many dictionaries and linguists are concerned, its place at the beginning of the UNICODE code chart made it an issue to be resolved.

**Issue # 2 [Level 3]**
The position of the *'hal'* sign (halant form) of a consonant in the sequence. Many alphabets of Sinhala do not explicitly specify the place of the 'pure consonant' form (the so called 'vowel removed form') of Sinhala letters. As such, there is common confusion as to its rightful place in the alphabetical order. For digital representation, this becomes an important issue to be resolved.

**Issue # 3 [Level 3]**
The positions of words containing *yansaya, rakransaya* and *rephaya* when there are two or more alternative forms available for the same word. While in general there is agreement that these 'short forms' are exactly equivalent to their non-shortened forms, in a digital representation a decision has to be forced as to which of them precedes the other.

**Issue # 4 [Level 2]**
Miscellaneous issues such as the archaic way of writing words such as **කාර්යාලය** and the irregular forms ක්‍රෑර or ක්‍රෘර and even කෲර. Though rare, the exact function and position of such words in a sorted list of words needs to be explicitly given to facilitate digital processing.

**Issue # 5 [Level 3]**
Whether 'ඥ' is a ligature of ඥ and ඤ or a single letter; and the right position of 'ඥ'. This again is an issue raised by its existence as a separate code point in the UNICODE code chart.

**Issue # 6 [Level 0]**
The position of the letter **ෆ**. The Latin symbol 'f' was superimposed on the Sinhala letter **ප** to produce the symbol **ඵ**, before the symbol **ෆ** was introduced into the alphabet. For this reason, and the phonetic closeness of the sounds represented by **ප** and **ෆ**, the letter **ෆ** has been popularly placed after the letter **ප** in many contexts. On the other hand, the letter **ෆ** being the newest letter of the Sinhala alphabet, is also placed at the end of the list of consonants in the alphabet.

**Issue # 7 [Level 1]**
There is a rule in the Sanskrit writing system that the consonant that comes after a *rephaya* is doubled, for example the words **චාතීතා**, **මාගීග**. The reason for this appears to be to display other vowel modifiers clearly when they are used with a consonant that comes after the *rephaya,* for example as in **කාම්මික**, **නාකීකික**, **ධම්මීම්න්මාදය**. The problem arising with

* N.B : *Issues # 6, 7, 8* were not taken into consideration in the first phase of the survey. The *Levels* indicate the perceived severity of the issue concerned, 3 being the most critical.

this kind of phenomenon when sorting is whether their positions should be considered based on this doubled form, or on their corresponding simplest form (as **කාර්මික**, **මාර්ග**, **තාර්කික**, **ධර්මෝන්මාදය** in the examples above).

**Issue # 8 [Level 1]**
Finally, the sort order of words which consist of intra-word spaces, for example, the name **ද සිල්වා**. The issue here is whether to consider this as two words, to ignore the space and consider the string as a single word, or to consider the entire string including the space as the single full word.

While these issues were the ones pre-identified in the study, the availability of online tools for testing any suggested collation sequence expressed in accordance with the UNICODE Collation Algorithm specification, allowed us to look for any other issues which might be 'thrown up'.

## 5. Observation made in Dictionaries

Some of the main and popular dictionaries were selected from among the various Sinhala dictionaries published for the purpose of this study. These dictionaries were selected after considering multiple attributes such as their quality, quantity in circulation, real usage and the perceived degree of authority of their compiler(s).

The following were the dictionaries selected for the present study:

[1] *Sri Sumangala Shabdakoshaya* compiled by Ven. Velivitiye Soratha Thero
[2] *Sinhala Shabdakoshaya* published by the Department of Cultural Affairs
[3] A *Sinhalese* – English Dictionary compiled by Rev. Charles Carter
[4] *Prayogika Shabdakoshaya* compiled by Dr. Harishchandra Wijetunge
[5] *Sinhala Vishvakoshaya* published by the Department of Cultural Affairs

The issues identified regarding the collation order of Sinhala were kept in mind while these dictionaries were being studied. The information gathered from these dictionaries regarding each issue is summarized in Table 1.

Based on the above, some partial conclusions could be made as follows:

*Issue 1:* This appears to be a non-issue as far as dictionary compilers are concerned. These two semi-consonants are placed at the end of the set of vowels in the Sinhala alphabet.

*Issue 2:* Apart from the *Sri Sumangala* dictionary, each of the other 4 had a clear decision that the 'hal' form comes after all other vowel derivatives. While the justification given by each of these two schools makes sense in their own contexts, the majority decision may need to be adopted for our purposes. Section 6 illustrates the difference between the two schemes.

*Issue 3:* Interestingly, none of the dictionaries are able to shed light on this issue owing to each only containing a single form – either the short or the non-short. As such this issue cannot be resolved using this methodology.

*Issue 4:* There is wide variation on the treatment of this issue. As such, a final decision on resolving this is deferred at this stage.

*Issue 5:* All dictionaries implicitly consider 'ඥ' as a ligature by their positioning of words beginning with it appearing soon after those beginning with 'ඥ'+ 'ඣ'.

*Issue 6:* In all dictionaries which includes it, the position of the letter **ඎ** is immediately after the consonants.

*Issue 7:* The doubling of the *reph*-modified consonant is given as a spelling variant of the simpler form in all dictionaries which contained it.

*Issue 8:* All dictionaries include words which have intra-word spaces where appropriate.

## 6. Procedures Followed by Dictionaries to Sort Words

The procedure followed in the *Sri Sumangala Sabdhakoshaya* to arrange words manually according to the alphabetical order is best specified by the following algorithm:

1. *Identify the syllabic units*[1] *of the two words*
   *Let the two words be* $w_1$ *and* $w_2$

[1] Syllabic unit means entity that contains a consonant and a vowel which is represented with *pilli.* E.g. කි,කො, කෑ are syllabic units and in some contexts ක, අ ,අා can also be syllabic units, in අක්කා අ and ක් are syllabic units. The syllabic units of අත්තික්කා are අ, ත්, ති, ක්, කා.

| Issue | Sri Sumangala Shabdakosaya | Sinhala Shabdakosaya | Carter's Sinhala English Dictionary | Prayogika Shabdakosaya | Sinhala Vishvakoshaya |
|---|---|---|---|---|---|
| The positions of the *'anusvara'* and *'visarga'* | In all the dictionaries the *'anusvara'* and *'visarga'* come at the end of the vowels. | | | | |
| The position of *'hal'* sign | The criteria followed is described in the Section 6 | A letter with *'hal'* sign comes **after** that letter's vowel derivatives. (The criteria followed in the dictionaries is described in the Section 6) | | | |
| The positions of the words containing *'yansaya'*, *'rakaransaya'* and *'repaya'*… | Observations could not be made regarding this issue since no dictionary uses two or more forms of the same word. | | | | |
| Irregular forms | ක්‍ර is used only to represent ක්රෑ not ක්‍රෑ | ක්‍ර=ක්රෑ ක්‍රෑ is is written as it is | ක්‍ර is used only to represent ක්රෑ not ක්‍රෑ | Both forms are followed in different places ක්‍ර=ක්රෑ or ක්‍රෑ | |
| Status of 'ඦ' | -considered as a ligature- | -considered as a ligature- | -considered as a ligature- | -considered as a ligature- | -considered as different char- |
| The letter ඏ | -not present- | ඏ comes at the end of consonants | -not observed- | ඏ comes at the end of consonants | ඏ comes at the end of consonants |
| Doubled *Reph* | -given along with the main entry as alternative spelling- | -given along with the main entry as alternative spelling- | -given along with the main entry as alternative spelling- | -given along with the main entry as alternative spelling- | -not observed- |
| Intra-word space | -space has been considered- | -space has been considered- | -space has been considered- | -space has been considered- | -space has been considered- |

**Table 1:** The position taken by Dictionaries on the eight issues under consideration

2. *Write each syllabic unit of both words as a consonant-vowel pair*[2]

3. *i=0*

4. *Choose the* $i^{th}$ *character of each word Let the two characters be* $w_1$ *(ch (i)) and* $w_2$ *(ch (i))*
   4.1 *If* $w_1$ *(ch (i)) = =* $w_2$ *(ch (i))*
      4.1.1 *i=i+1*
      4.1.2 *go to 4*

   4.2 *Else if* $w_1$ *(ch (i)) >* $w_2$ *(ch (i))*
      4.2.1 $w_1 > w_2$
      4.2.2 *break*
   4.3 *Else*
      4.3.1 $w_1 < w_2$
      4.3.2 *break*

Dictionaries other than *Sri Sumangala Sabdhakoshaya* compare consonant-vowel pair in a different manner which makes the two approaches different. In this method when two consonant-vowel pairs are compared two consonants and two vowels are compared separately. In the cases which vowels are not present the consonant of the next consonant-vowel pair is not taken as in the *Sri Sumangala Sabdhakoshaya*.

## 7. Views of Scholars/Academics and Linguists

The following scholars and academics were consulted with a view to acquiring their expert views – often based on their respective linguistic persuasions. The aim of the consultation was to attempt to achieve consensus and not just for documenting their independent views.

[2] If the syllabic unit does not consist of a vowel write 'null' in place of vowel (e.g. ක = ක null). The 'null' is considered as a character and it is the character that has the greatest value in the weight space.

- Professor Vinee Vitharana (VTH)
  *Chief Editor of the Sinhala Dictionary. Former Professor of Sinhala at the University of Ruhuna.*
- Professor Wimal G. Balagalle (WBA)
  *Former Chief Editor of the Sinhala Dictionary. Emeritus Professor of the University of Sri Jayewardenapura*
- Professor W. S. Karunathilake (WSK)
  *Former Professor of Linguistics at the University of Kelaniya*
- Professor J.B. Dissanyaka (JBD)
  *Emeritus Professor of the University of Colombo*
- Professor Sucharitha Gamlath (SGA)
  *Former Professor of Sinhala at the University of Ruhuna.*
- Dr. Harishchandra Wijethunge (HWI)
  *Author of the Prayogika Sinhala Sabdhakoshaya*
- Mr. Rupasinghe Perera (RUP)
  *Deputy Director, Pirivena Education Branch, Ministry of Education*
  *Secretary, Sri Lanka Oriental Languages Society*

Having explained the aims and objectives of this study, a list of lexemes that concretely represents all the possible issues was carefully designed and given to each consultant – rather than posing the issues in their abstract form. This approach forced an explicit response rather than inviting rigorous expositions of the theoretical basis for same. There were some issues which some of the linguists could not provide a direct answer to. However, most were able to make their suggestions as to how to resolve such issues by relying on their own linguistic theories. The books written by some of these scholars were also considered during this study. The comments made by each expert regarding the identified issues and their suggestions are summarized in table 2.

*Issues 1, 2, 6, 7 and 8* were not disputed by any of the experts who agreed with the majority (4) mainstream dictionaries. While *Issue 3* had no consensus solution, all experts agreed that there should be a single well-specified standard. *Issue 4* too had no clear consensus except for the recommendation that 'ක්රෝ' should be written as it is. There also appears to be a majority view that 'ඥ' should be treated as a ligature with the only dissenting scholar too later arriving at consensus in the interest of arriving at a consensus.

## 8. Procedures followed in State Institutions and Organizations

The following government organizations and institutes were selected for the purpose of identifying the different collation orders adopted by them for their regular work.

- National Library & Documentation Centre (NLDC)
  a. An explicit alphabetical order is available at NLDC
  b. The Sri Lanka National Bibliography is prepared according this alphabetical order
- National Institute of Education (NIE)
  a. The NIE has adopted the alphabetical order given in the *Sri Sumangala Shabdakoshaya*.
  b. This order is followed when school text books and recommended books for school children are prepared.
  c. The specified alphabetical order for government examinations (e.g.: GCE (O/L) and GCE (A/L)) is also the same.
  d. Further recommendations of the NIE are given in a separate publication titled *Sinhala Lekhana Reethiya*.
- Public Library – Colombo (PUB)
  a. The alphabetical order given in the *Sinhala Encyclopedia* is followed.
- Sinhala Dictionary Office (SDO)
  a. The criteria followed by the SDO is the criteria followed in the *Sinhala Sabdhakoshaya*
- Sinhala Encyclopedia Office (SEO)
  a. – not yet responded –
- Election Commissioner's Office (ELE)
  a. An explicit alphabetical order is available at ELE
  b. 'anusvara' and 'visargaya' comes at the end of vowels
  c. 'hal' sign comes at the beginning of vowels
  d. When there are two or more alternative forms available, the collation order is found according to the simplest form and the priority is given to the simplest form (issue #3)
  e. The letter 'ඥ' is considered as the conjunction of ඤ and ඣ.
  f. The recommendations of the NIE given in the *Sinhala Lekhana Reethiya* are also followed by the ELE.

| Issue | VTH | WBA | WSK | JBD | SGA | HWI | | RUP |
|---|---|---|---|---|---|---|---|---|
| 1 | At the end of the vowels | | | | | | | |
| 2 | At the end of the vowels | | | | | | | |
| 3 | Need to have a policy: what is simple should come first | Need to have a policy: whatever non-confusing | Need to have a policy | Need to have a policy: what is simple should come first | Need to have a policy: what is non-confusing | Need to have a policy: priority must be given to what is commonly written | | Need to have a policy: priority should be given to the tradition |
| 4 | ක්‍රෑ=ක්රෑ ක්රෑ should be written as it is | ක්‍රෑ=ක්රෑ ක්රෑ should be written as it is | Sanskrit loan words should be written in their traditional forms, but English loan words can be written in either form | ක්‍රෑ= ක්රෑ කෘෂ= ක්රෑ | ක්‍රෑ=ක්රෑ ක්රෑ should be written as it is | ක්‍රෑ=ක්රෑ ක්‍රෑ= ක්රෑ | | ක්‍රෑ=ක්රෑ ක්රෑ should be written as it is |
| 5 | 'ඥ' is a ligature ඥ+ ඤ | 'ඥ' is a ligature ඥ+ ඤ | 'ඥ' is a ligature ඥ+ ඤ | 'ඥ' is a single letter | 'ඥ' is a ligature ඥ+ ඤ | 'ඥ' is a ligature ඥ+ ඤ | | 'ඥ' is a ligature ඥ+ ඤ |
| 6 | ් comes at the end of consonants | ් comes at the end of consonants | ් comes at the end of consonants | ් comes at the end of consonants | ් comes at the end of consonants | ් comes at the end of consonants | | ් comes at the end of consonants |
| 7 | -give along with the simplest form- | -give along with the simplest form- | -give along with the simplest form- | -give along with the simplest form- | -give along with the simplest form- | -give along with the simplest form- | | -give along with the simplest form- |
| 8 | -space has to be considered- | -space has to be considered- | -space has to be considered- | -space has to be considered- | -space has to be considered- | -space has to be considered- | | -space has to be considered- |

**Table 2:** The position taken by Linguists on the eight issues under consideration

- Library – University of Colombo (UOC)
  a. The alphabetical order given in the *Sinhala Sabdakoshaya* is followed.
- Library- University of Peradeniya (PDN)
  a. The alphabetical order is the same as that used by UOC.
- Library- University of Kelaniya (KLN)
  a. – not yet responded –
- Library- University of Sri Jayewardenepura (SJP)
  a. –not yet responded –
- Library-University of Ruhuna (RHU)
  a. The alphabetical order given in the *Sinhala Encyclopedia* is followed.

Of the above, the NIE and Election Commissioner's Office (ELE) deserve special attention. The ELE standard is of interest to this study because it explicitly addresses the issues at hand – Issues 1 through 3

and 5. In Issues 1 and 2, the ELE standard tallies with those of the majority dictionaries and linguists. Interestingly, ELE has a definite recommendation for Issue 3 – i.e. to locate all such form variations together at the rightful place of the *simplest* form with the simplest form preceding the other forms in decreasing order of simplicity. Finally on *Issue 5 (6, 7 and 8)* too, the ELE standard concurs with that of the expert consensus.

Since the government recognizes the NIE as the prime authority in setting educational standards the order recommended by them becomes of utmost importance. Some of the other reasons for attaching such importance to this recommendation include:

(a) The standard specified has been created by a representative groups of scholars and linguists including many of those consulted in the present study.
(b) Generations of school teachers and students have already adopted this standard and hence it is the closest to a *defacto* standard.
(c) Their more recent publication, *Sinhala Lekhana Reethiya*, is widely used by state organizations including the Commissioner of Elections.

## 9. Summary Recommendations

The status of each of the issues considered in this study together with the recommended solution is presented below.

*Issue 1:* The dictionary survey and ratified by the expert consultation resolved this issue to the satisfaction of the authors: treat both the *'anusvara'* and *'visarga'* as appearing in the alphabetical order immediately after all the vowels. This is also further confirmed by the ELE and NIE standards which are in wide practical use.

*Issue 2:* The dictionary disparity with regard to the correct position for the 'hal' form was resolved by a unanimous opinion by the experts consulted that it should immediately follow the vowels but precede the *'anusvara'* and *'visarga'*.

*Issue 3:* This was one of the issues on which empirical evidence was scarce. However, the openness of all the linguists for *some* standard and the *simplicity* rule recommended by some of them and clearly enshrined in the ELE standard is to order all forms of such words adjacent to each other beginning with the simplest form and increasing in complexity. This would prescribe an the following order on the three common forms of the work *karyalaya*: **කාර්යාලය**, **කාර්‍යාලය**, **කාර්‍යාලය**.

*Issue 4:* This is the issue with the greatest degree of divergence in opinion. Three of the dictionaries and five of the linguists however concurred that 'ක්‍ර is used only to represent ක්‍රෑ not ක්‍රෘ'. The latter is represented as it is. This is in contrast to the original Sinhala UNICODE recommendation where 'ක්‍ර= ක්‍රෑ and ක්‍ර= ක්‍රෘ'. It seems prudent to adopt the majority opinion.

*Issue 5:* This was the single main success in the consensus seeking process. It is thus recommended that 'ඳ be treated as the ligature of ඒ+ ඳ' so that it does not appear in the order thought to be implied in the UNICODE code chart.

*Issue 6:* This seemed to be an issue as it was introduced later to the Sinhala alphabet and the phonetic similarity of the letters **ඵ** and **ඟ**. There was confusion with the symbol **ඞ** too. According to *Sinhala Lekhana Reethiya* , the book published by the NIE for Sinhala, and all the scholars it is accepted that the letter **ඟ** should come at the end of the consonants.

*Issue 7:* This form is used merely for representation purposes. The underlying meaning of both **කාම්මික** and **කාර්මික** are the same. Therefore they occupy the same collation position. In dictionaries these are given along with the main entry as spelling variations.

*Issue 8:* It is important to consider intra-word space when sorting is done in some domains (e.g. directories of names). However, this cannot be prescribed in the alphabet but at the level of the particular application.

Based on the above recommendations and extensive testing done using early versions of the proposed collation sequence, a UNICODE Collation Element table together with its weights is recommended as the explicit specification of the Sinhala alphabet for use in electronic processing of Sinhala. The documents observed at the organization mentioned above and other relevant documents including the proposed Collation Element Table can be found at the URL http://www.ucsc.cmb.ac.lk/ltrl/public/collationDocs.

html.

## 10. Conclusion

At the outset we pointed out that a complete and unambiguous specification of the Sinhala alphabet is an essential and urgent requirement for all kinds of electronic processing of Sinhala text. The process of study revealed five major areas unresolved as far as the Sinhala collation order was concerned and three other areas which needed clarification. We outlined a methodology of arriving at a set of well informed recommendations based on three sources: widely accepted dictionaries, the most respected Sinhala scholars and the most widely adopted official standards on collation sequence.

Using a consensus-based approach, we have successfully arrived at a unique collation sequence for the Sinhala language and expressed it explicitly using the UNICODE Collation Algorithm specification of the UNICODE Consortium. Testing of this specification for arbitrary lists of words is made possible by online tools available from ICU.

***Acknowledgements***

*This work was partially supported by IDRC under the PAN Localization Project and sponsored by the ICT Agency of Sri Lanka. The authors are indebted to all Sinhala Language scholars, Publishers and other Practitioners who collaborated willingly in the work described in this study. In particular we wish to acknowledge the contribution of Mr. Harsha Wijayawardhana and Mr. Asanka Wasala as well as other colleagues in the Language Technology Research Centre of the University of Colombo School of Computing.*

## References

Karunathilake, W.S , *Sinhala Basha Viyakaranaya*, M.D.Gunasena & Company Ltd, 1997.

Wijetunge,H , *Sinhala Akuru Akaradhie Kireema Pramitha Kireema*, S. Godage Brothers, 2003

Perera, R, *Sinhala Vahara Athpotha*, Thivira Publishers, 2004

*Sinhala Lekhana Reethiya*, National Institute of Education, 2001

Disanayaka, J.B, *Akuru ha Pilli*, S. Godage Brothers, 2000

# Using Web Services for Translation

## A white paper on the Translation Web Services Standard

**Kevin Bargary[1], Peter Reynolds [2]**
**[1] Localisation Research Centre, University of Limerick**
**[2] Lionbridge, Dublin**
**Ireland**
**kevin.bargary@ul.ie, peter.reynolds@lionbridge.com**

**Abstract**

Web services use Internet technologies to allow computer-based systems to communicate and transfer data in a way that provides a more seamless automated workflow. At OASIS work is being done to create a standard way for Web services to be used within the translation and localisation industry. The purpose of this article is to inform you about this work, what Web services are, and to outline a real-life case study showing how this technology is being put to use now. The article will deal in detail with the specification proposed by the OASIS technical committee for Translation Web Services. It will also describe use cases where this specification can be put into practise such as the project implemented by the Localisation Research Centre as part of the IGNITE project.

**Keywords:** *Open standards, Web services, translation, localisation life cycle, standards development, OASIS, localisation, localization, globalisation, globalization, Localisation Research Centre, LRC*

### 1. Introduction

Web services is a solution to the problem of computer systems not talking to each other. It uses Internet technologies to allow computer-based systems to communicate and transfer data in a way that provides a more seamless automated workflow. The likelihood is that Web services will be adopted by companies over the next few years to automate processes and integrate systems. Within the translation and localisation industry, however, there is work being done to create a standard way for Web services to be used. This work is being done at OASIS, which is the organisation for creating standards, particularly XML standards within the industry. The purpose of this article is to inform you about this work, what Web services is, and to outline a real-life case study showing how this technology is being put to use now.

The idea of creating a standard for the use of Web services within translation was put forward by Bill Looby of IBM at the eLocalisation 2001 conference held in Limerick, Ireland. Mr. Looby delivered a paper which showed a vision of how the industry could benefit from agreeing on a common way of using this technology. The conference had also seen a practical demonstration of how this technology was already being used. Lionbridge (then Berlitz GlobalNET) gave a demonstration of work being done for the 2003 Special Olympics Web site, which it was sponsoring. Using Web services, an XLIFF file was sent from the Web site to Elcano, Lionbridge's online translation service, and back to the Web site.

This conference ended with a small group of people getting together to look at how they could progress with the idea of using Web services within the translation industry. The steering group that formed decided that OASIS would be the natural home. OASIS was established in 1993 and it is focussed on XML standards for the software industry. XLIFF (XML Localization Interchange File Format) was already being developed by an OASIS technical committee, and there was considerable support for this industry within OASIS. The OASIS technical committee was formed at the beginning of 2003 and its members include representatives from Oracle, Microsoft, IBM, Connect Global Solutions, thebigword, LISA, the LRC, and Lionbridge as well as individual members.

### 2. Translation Web Services

#### 2.1 What are Web Services?

Before detailing the main features in the draft specification from the Translation Web Services (TWS) technical committee, we would like to give some background on Web services. The World Wide Web is a collection of interlinked documents that sits on the Internet, which is effectively a huge computer network that connects individual Web sites. To access a Web site a person sits at a computer and views pages through a Web browser — a process that can be considered as machine-to-person communication. With the advent of Web services the Internet is used for machine-to-machine communication rather than

machine-to-person communication. Protocols such as HTTP and standards such as XML and SOAP (Simple Object Access Protocol) are used in Web services to enable this machine-to-machine communication. This enables different systems to work together, allowing for more powerful functionality and automation.

A Web service might do something such as enable weather forecasts to be queried by remote computers over the Internet. To do this the weather forecasting company would have to create a Web service and allow it to be accessed. This is done using an XML document called a Web Service Definition Language (WSDL) document. The WSDL document describes the services which the client application is allowed access to and describes what parameters will be sent and received for each of these calls. A gardening enthusiast who is away a lot might want his computer to control his water sprinkler. By using Web services the computer will be able to find out when it is not raining and turn on the sprinkler. A protocol called Simple Object Access Protocol (SOAP) is used for this.

## 2.2 What is Translation Web Services

Since January 2003 the Translation Web Services (TWS) technical committee has been working to create a standard way for Web services to be used in a multilingual context. It has concentrated its efforts on creating a standard relating to the communication between publisher and vendor companies. At the simplest level this will allow for translation and other work to be sent by the publisher to the vendor and, once translated, sent back. The draft specification covers the following areas:

- Service support
- Translation and request quote
- Status, notification and delivery
- Reference files
- Security

## 3. Translation Web Services Specification

Each service in the TWS specification provides two forms for interaction between client and vendor. These forms are *request* and *response*. For example, the client submits a *retrieveServiceList* request to the vendor. The vendor receives this request, processes it, and then returns a *retrieveServiceList* response to the client. The *request* and *response* forms of a service expect different inputs and produce different outputs but both *request* and *response* are needed for the interaction between and use of each service.

### *3.1 Categories of Service*

The TWS specification defines five categories of methods or services, namely 'Service Support', 'Security', 'Translation & Request Quote', 'Status, Notification and Delivery' and 'Reference Files'. These categories form a guideline for the services that the TWS specification provides. Each category encapsulates one facet of the core work required for the completion of a translation job, from initial quote through to final delivery.

#### *3.1.1 Service Support*

The 'Service Support' category contains only one service, namely *retrieveServiceList*. This service allows the client to query the vendor on the type of localisation services they provide. When a *retrieveServiceList* request is made on a vendor a list of languages, service types, domain types, and MIME (Multipurpose Internet Mail Extensions) types that are supported is returned. In the case where no relationship exists between client and vendor, the *retrieveServiceList* is the first service evoked by the client to ensure the vendor meets the requirements for the potential translation job. The client can then use the information returned in their future interactions with the vendor.

#### *3.1.2 Security*

As with any transaction over the web, data security is an important consideration. OASIS defines a Web services security standard specification (WS-Security) which provides several methods for the securing of Web service-related transactions. The TWS specification relies on WS-Security to provide an end-to-end message level security and hence the specification recommends the use of username/password-based security over SSL.

#### *3.1.3 Translation and Request Quote*

This category details the services required to instantiate a job between a client and a vendor. Web services for translation uses a job ticket as a unique identifier for each project. This job ticket is created on the client side usually before a quote request. The job ticket consists of a project ID, a user ID and a unique job ID. This job ticket can then be used in all future interactions with the vendor's web service. Currently there are two methods of initiating a job in the 'Translation and Request Quote' category.

The first method is where the client submits a

*requestQuote* service. The *requestQuote* service details the information pertaining to the translation job (word count, languages etc.). The client retrieves the generated quote using the service *retrieveQuote* and chooses to accept or reject the quote. If the quote is accepted an *acceptQuote* service is activated, if not accepted the generated quote will expire after a certain time limit, defined by the vendor.

The second method is based upon the *submitJob* service. The *submitJob* service has similar inputs to the *requestQuote* service but it also contains the purchase order information found in the *acceptQuote* service used in the previous method. In using this second method it is automatically assumed that the job will be accepted. This interaction might be between two in-house systems, e.g. one system has content to be translated, and it contacts a second MT system and gets the content translated.

### *3.1.4 Status, Notification and Delivery*

The Translation Web Services technical committee provides seven status, notification and delivery management services in the TWS specification. This set of services allows the client some control over the work that is being carried out by the vendor for a particular job. Using these services a client can check the status of and cancel or suspend a particular job. The success of each service request is dependant on the state of the job at the time of calling the service. For example you cannot cancel a job that has already been completed (for obvious reasons).

The *retrieveActiveJobsList* service returns to the client a list of all *active* jobs that they have with a particular vendor. An alternative to this is the *retrieveFullJobsList* service, which returns *all* jobs associated with a particular vendor irrespective of the current status.

A client can query the vendor using the *retrieveJobInformation* service and get a response containing all current information about a job. The status of the job can be deduced from the information received from the *retrieveJobInformation* service and possible changes to the project deadlines can be predicted. If a job is completed then the status of the *retrieveJobInformation* service response should reflect this.

If the information received back from the vendor after a *retrieveJobInformation* service request indicates that the job is completed, this job can then be downloaded using the *retrieveJob* service.

The client can choose to suspend a job temporarily at any time as long as the job status is not complete. This is done by making a *suspendJob* service request.

To remove this temporary suspension of a job (by submitting a *suspendJob* service request), the client can choose to resume the job using the '*resumeJob*' service.

A client can cancel a job using the *cancelJob* service if the job is currently active and not in a completed state.

### *3.1.5 Reference Files*

The vast majority of localisation projects require not only the localisable content but also any reference files associated with the project. Reference files are not for translation but contain information that may help the translation process. Translation memories, style guides, or terminology references may be sent along with the translatable files. The 'Reference' category defines services to allow for this allocation of these files to a particular project.

A resource file can be assigned to any number of active jobs using the *associateResource* service. This service can also be used in batch mode which allows users to upload numerous resources at once. There is also a 'description' field that indicates what a particular file is for.

To remove an association between a job and a resource file the *disassociateResource* service is used.

The client can review information about a resource file by invoking the *retrieveResourceInformation* service. This will return information about the resource file from the vendor: a list of jobs that the resource file is assigned to, the purpose of the resource file and whether or not the file has changed (been updated).

The TWS specification allows the client the functionality of uploading assets to the vendor using SOAP messages using the *uploadFile* service. This *uploadFile* service also can accept multiple file uploads in a batch and has an identifier field for the files being uploaded for easier recognition of the files.

## 3.2 Services Supported in Current Specification

At this point there are 18 services being supported by the TWS specification. As discussed in section 3.1, services can be categorised into one of five categories depending on their function in a translation process. These services can conversely be considered under the following three headings:

**Required Services** — these are services that are required for a Translation Web Services implementation and form the basis of a minimalist approach to Translation Web Services use.

**Recommended Services** — these services are recommended by the Translation Web Services technical committee to be used in an implementation of Translation Web Services together with the 'required' services.

**Optional Services** — these services are services that are only needed in some specific cases (enquiring about what services a vendor has to offer or setting up a first contact with a vendor by requesting a quote etc.). These services are not essential to the Translation Web Services process but are required for some scenarios.

| Required Services | Optional Services | Recommended Services |
|---|---|---|
| submitJob | retrieveServiceList | rejectJob |
| retrieveJobInformation | requestQuote | associateResource |
| retrieveJob | acceptQuote | disassociateResource |
| retrieveActiveJobsList | retrieveQuote | retrieveResourceInformation |
| suspendJob | retrieveFullJobsList | retrieveFullResourceList |
| resumeJob | | uploadFile |
| cancelJob | | |

**Table 1:** List of services available in the TWS specification

## 4. Technologies in Translation Web Services

### 4.1 Simple Object Access Protocol (SOAP)

SOAP is a W3C-developed standard described as a communication protocol or a message passing system between two computers. SOAP is the specification that defines the XML format for these messages being passed. SOAP is one of three core XML-based standards that are the foundation of a Web services implementation (the others being WSDL and UDDI, see Section 4.2 and Section 4.3).

### 4.2 Web Services Description Language (WSDL)

A WSDL is an XML document that describes (a) a set of SOAP messages and (b) how these messages are exchanged. The WSDL file in a practical sense contains the information required by a client to access a service, i.e. what parameters need to be passed to the service to invoke a response. The WSDL file should also contain the actual location (HTTP address) of the web service.

### 4.3 Universal Description, Discovery and Integration (UDDI)

UDDI is an OASIS-driven mechanism for clients to dynamically find other Web services. The UDDI protocol gives a company the ability to register their available Web services online, thus exposing them to potential clients. An UDDI registry service is a Web service that manages information about service providers, service implementations, and service metadata.

In summation, SOAP is the communication protocol for Web services, WSDL defines how the interaction occurs between the two computers, i.e. how to invoke the services and the UDDI is a mechanism for finding these services or registering one's own services.

## 5. Localisation Life Cycle

See figure 1: Localisation life cycle incorporating the use of Web services for translation

## 6. Use Cases

### 6.1 TWS Reference Implementation

A reference implementation of the Translation Web Services specification was undertaken by the Localisation Research Centre at the University of Limerick in Ireland as part of the IGNITE project. The Translation Web Services technical committee

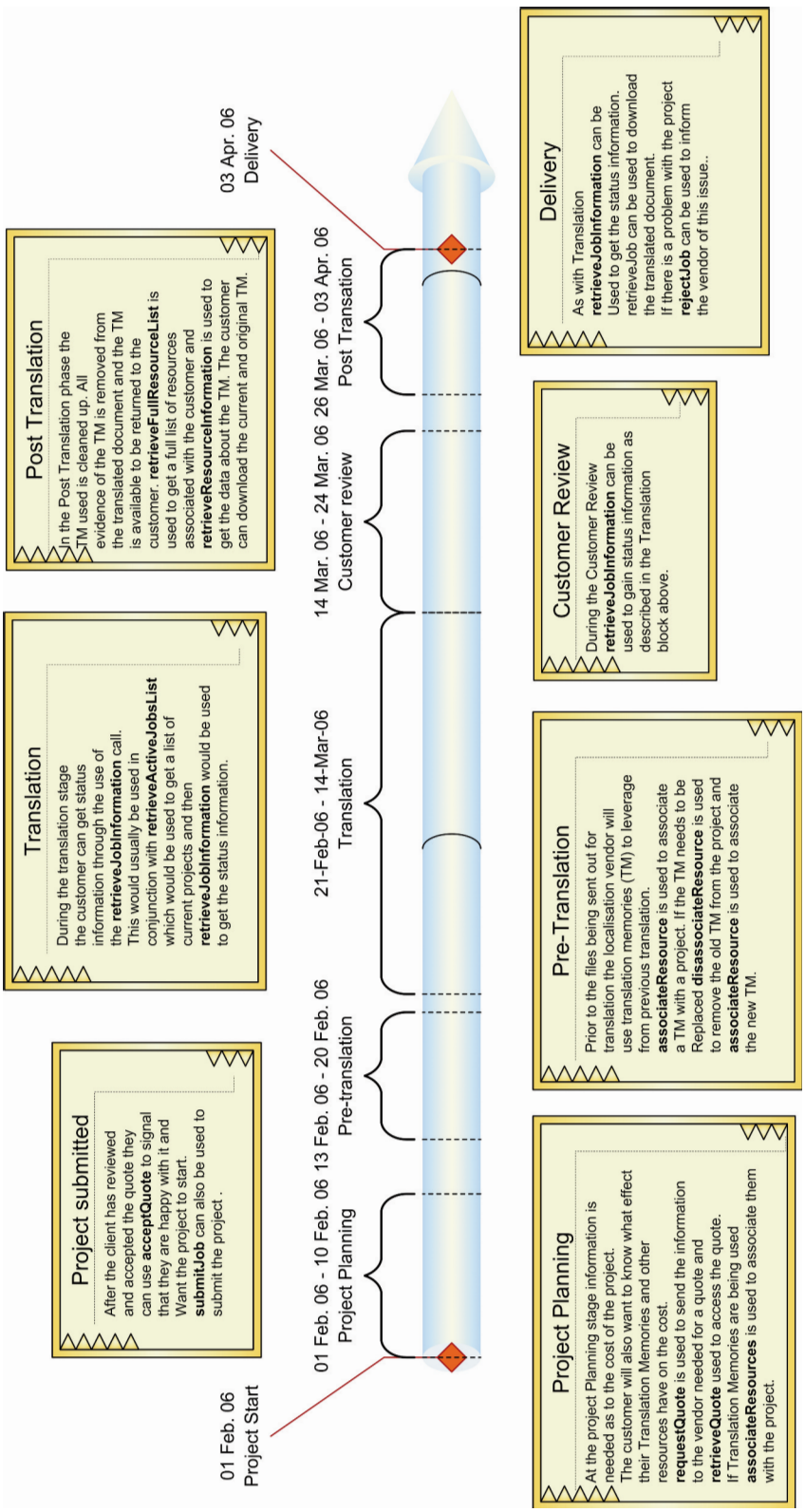



**Figure 1**: Localisation life cycle incorporating the use of Web services for translation

decided that it was important to have a reference implementation to see how the standard worked from a technical viewpoint.

The basic premise of Web services for translation is that a server machine will contain a pre-programmed set of methods or functions that a client machine can access using Web services technology. The two components involved in this interaction are the server machine and the client machine. The process of implementing Web services for each component is similar. Nevertheless there are some subtle but important distinctions to be made between both implementations.

The implementation platform of choice for this reference implementation is J2EE and the Java programming language. The rationale behind this decision was that the open source 'Apache Project' has a Java-based implementation of SOAP called *Axis*. Axis is defined as a "reliable and stable base on which to implement Java Web services"; it provides an Application Programming Interface (API) into the SOAP actions that are required for implementing the Translation Web Services standard. Apache Tomcat was chosen as the Web server on which to develop the Web services. Tomcat and Axis both developed under Apache work very well together in a practical implementation environment.

The development model used for this reference implementation was loosely based on the prototyping model of software development. Initially one of the 18 services currently available in the standard was developed. From that the development incorporated one further service at a time until all were implemented. Throughout the development life cycle we reported back to the technical committee on any issues or suggestions for improvements that arose as we progressed.

Apart from the API for the usage of SOAP functionality Axis provides two command line utilities that are essential to the implementation of Web services. The 'Java2WSDL' utility takes pre-existing Java code and creates a WSDL file for that code. This utility can be used if there is some code that performs a specific function that you would like to make available as a Web service for others to use. The Translation Web Services standard has made a WSDL available, so this utility was redundant for our purposes. However, the second utility, 'WSDL2Java', creates the Java stubs (Java files that contain the code needed to use SOAP) required by:

(a) the server to write and deploy the service, and
(b) the client to access the service through its own code.

Figures 2 and 3 show this process. Firstly the Java stubs are created and then the Java stubs are used by the client application to access the services and by the server to deploy the services.

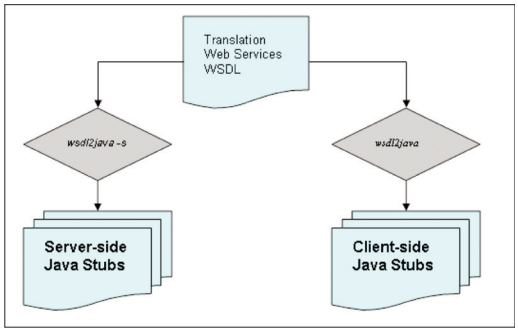



**Figure 1:** Creating the Java Stubs from the WSDL

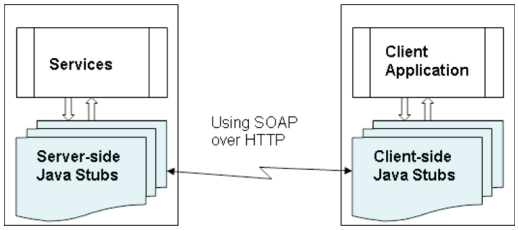



**Figure 3:** Connecting client and server (through the Java stubs) over HTTP using SOAP

The first service that we implemented was the retrieveServiceList service. This service was chosen because there were no input parameters required for it. All that was required to invoke the service was an instance of the *retrieveServiceListRequest* class. The *retrieveServiceList* service returns "a complete list of services offered by a particular vendor. This will include the languages dealt with and services offered by a particular vendor" (Translation Web Services Specification Draft 1.0). After writing the server-side code to handle a *retrieveServiceListRequest*, i.e. return all of the appropriate values, the next stage was to create a simple test class that could instantiate a *retrieveServiceListRequest* and handle the results received back from the server in a *retrieveServiceListResponse*. With both classes now ready, we needed to deploy the services to the Apache Web server. The WSDL file also contains the location of the service, i.e. where it can be accessed from.

Deployment is necessary to ensure the service is in the location as defined in the WSDL.

When the utility 'WSDL2Java' creates the Java stubs needed for the server-side machine, it also creates two other files that are used to deploy and 'un-deploy' the service to a Web server (Apache Tomcat). These files are called Web Service Deployment Descriptors ('deploy.wsdd' and 'undeploy.wsdd').

The next stage in the development of the implementation was to write the code for the rest of the services. While writing the code we encountered some issues with the specification (including inconsistencies between the schema and the specification document). These issues were quickly amended by the technical committee. During the process of coding we also made some suggestions to the technical committee about possible improvements to the specification and we were actively involved in applying these changes. For example, the service '*retrieveQuote*' in the original specification did not return any information about the location of the actual quote. This was deemed to be an important piece of information for this service and was promptly included in the specification.

With the code for the implementation of the services now written, the initial service deployed (*retrieveServiceList*) was un-deployed and the full list of services was deployed to the Web server. A JSP (Java Server Pages) client interface was developed to allow for the input of the parameters required for each service and also to show the responses from the server.

### 6.2 Next Steps

The next step in the reference implementation will be to attach the front-end JSP interface to a back-end database system that returns some relevant information that is not pre-defined. The current implementation can be seen running live on www.electonline.org:8080/index.html. To view or download the source code used, please go to www.igniteweb.org/tws. Here you will also find instructions on how to install this implementation on your own local machine and server.

### 6.3 Lionbridge's use of Web services

Although the specification from the TWS technical committee is still at the draft stage, there has been some significant work done with Web services in the translation industry. Lionbridge has implemented a number of solutions based on Web services which have linked content management and other systems with Elcano, its online translation portal.

**Freeway**

Lionbridge recently introduced its new customer portal in April 2006 and the Elcano Web services described above will be ported to Freeway. More information is available at www.lionbridge.com.

***Notes***

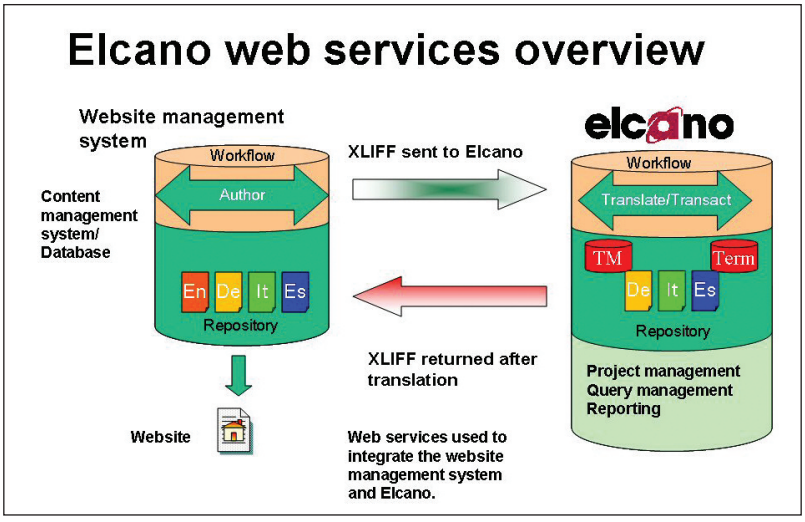



**Figure 4:** Lionbridge's use of Web services

*This article was written by Kevin Bargary and Peter Reynolds with additional contributions from Magnus Martikainen, Tony Jewtushenko, Andrzej Zydron and Reinhard Schäler.*

***References***

*Not all of the websites listed below are still active. Links are only provided for the purpose of presenting an accurate reproduction of the original paper.*

Translation Web Services Technical Committee (2006) [online] http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=trans-ws

XLIFF Technical Committee (2006) [online] www.xliff.org

OASIS (2006) [online] http://www.oasis-open.org/

World Wide Web Consortium (2006) [online] www.w3.org

Localisation Research Centre (2006) [online] http://www.localisation.ie/

IGNITE project(2006) [offline] http://www.igniteweb.org/

LISA – OSCAR (2006) [offline] Available at http://www.lisa.org/sigs/oscar/

# Formatting and the Translator: Why XLIFF Does Matter

**Ignacio Garcia**
**University of Western Sydney**
**Australia**
**i.Garcia@uws.edu.au**

Gains in productivity through translation memory-based text reuse are often offset by time spent in dealing with formatting glitches. This affects all players in the localisation industry, from the end client to the language vendor to the freelance translator. However, as a non-core activity for them, translators are less well prepared to deal with these hidden formatting related costs. This article looks from the translator's viewpoint at the importance of formatting as part of the translator's work, and at the limitations in dealing with formatting of the technologies now in use. It also shows how the development and implementation of standards within the localisation industry, XLIFF in particular, may impact on the situation, so that translators can once again deal only with text, as they did in pre-digital times.

**Keywords:** *Language Industry, Localisation, Open Standards, Text Reuse, TMX, XLIFF*

## 1. Translators translate files, not text

What translators receive for translating is files, not just text. Translators do not receive TXT files, but files with text plus formatting; with data that users can read plus code that machines can. Since many of the files translators receive have been formatted in Word, which we are all familiar with as the *de facto* standard for word processing, some may assume that formatting is transparent and has nothing to do with translation. However, the fact that translators, as computer users, do not need to 'read' the code to understand the text does not mean they don't need to pay attention to it. Translators who have been exposed to other formats have learnt that it pays to understand the differences between flat and binary files, and between structured formatting and inline formatting. The digital world has created both the file, an amazing advance from the days when text was composed on a typewriter, as well as specific technologies to deal with translation, principally translation memory (TM). This digital world has also raised the issue of formatting. It is argued here that gains in productivity through TM-based text reuse are often offset by time spent in dealing with formatting glitches.

That formatting is part of the translator's job is obvious to any translator working for the localisation industry. Formatting, however, has not been given the prominence it deserves in training and professional development. There is no mention of it in the Language Engineering for Translators Curricula (LETRAC) Curriculum Modules (1999) in which many of the programmes with a focus on technology were first based. Even today those programmes tend to present formatting as something that will be taken care of by specialised computer software, TM suites or localisation tools. This is not quite the case yet. The importance of formatting, notwithstanding the technology currently available, has been repeatedly pointed out in the literature addressed to language vendors and end clients (Reynolds & Jewtushenko, 2005). There is a gap, however, in the literature addressed to the freelance and the trainee translator that this article will attempt to fill. Austermuhl (2001) hardly refers to formatting; Bowker (2002: 37-39, 118–119) and Somers (2003: 18–19) only treat it marginally. Only Zetzsche (2003b) pays thorough attention to it, its focus being to give the freelancer the tools to deal with digital text.

To some extent, it is understandable that not much profile is given to formatting in academic settings. Text is the core issue for translators, formatting is not. Dealing with formatting, like dealing with invoicing, may be a most important activity, but it is non-core. Also, translating text is a complex activity that takes years to master. It involves weighing up alternative renditions of a meaning in the target language in order to choose the most appropriate one for the situation, in a context where there is rarely a clear-cut right or wrong answer. Dealing with formatting, on the other hand, may be very complicated, but those who manipulate files will realise soon enough whether or not they have done the right thing. It is, however, part of the translator's job, as current technology is not yet good at separating text from formatting (i.e. content from its container) within the file. In the age of the typewriter and before, formatting was unimportant. In

the first stages of the digital age, it has become important, and it will continue to be important — at least until we reach the 100% XML scenario outlined below.

## 2. Translation is not a craft — it is an industry

Translation is no longer a craft; it is an industry. However, it is an industry which does not pay the translator — the freelance at least — by the year or even by the hour like respectable professions such as law and medicine do, but by the word (or by the line, or by the page: by quantity). Translators work at the 'wordface' in the same way that miners work at the coalface, as Emma Wagner put it (Chesterman & Wagner, 2002: 1), taking out 'loads of translated words' which is what language vendors sell, as Mark Lancaster, the head of SDL, a major language vendor and the most important developer of computer-aided translation tools, was reported to have said (in Fenstermacher, 2006). On the one hand, there is a low threshold entry point to the profession: any educated bilingual, given enough time and some mentoring, can become a translator; on the other, only those able to translate at great speed will be able to make it professionally profitable.

Most translators work within what has been loosely called the language industry or, more precisely, the localisation industry, also referred to lately as the globalisation industry, or the GILT (globalisation, internationalisation, localisation and translation) industry. This is an industry that, whatever name it uses, is based on selling lots of translated words, with quality often taken for granted, time-to-market an important constraint, and price, paramount. This is an industry that, according to the latest calculations and with conservative estimates, will be worth more than 9 billion US dollars by the end of 2006 and will grow at 7.5 percent per year to be worth an estimated excess of 12 billion US dollars by 2010 (DePalma & Beninatto, 2006: 4-5). Language vendors, like individual translators, are also paid by the number of translated words they deliver to the end client, with a benefit margin that can only be widened through increases in productivity. Despite efforts by the industry itself to monitor quality (the Localization Industry Standards Association (LISA) being a case in point) and initiatives such as the recent EN-15038 European Quality Standard for Translation Services, backed by the European Committee for Standardisation (Arevadillo Doval, 2005), the translation industry also has a low entry threshold and does not require a large amount of capital. So there is fierce competition, competition that shows in a consolidation process best reflected at the top end of the market in the mergers and acquisitions of Mendez by Lernaut&Hauspie, then of Lernaut&Hauspie and Berlitz by Bowne, then of Bowne by Lionbridge, a process that does not seem to have stabilised yet (DePalma & Beninatto, 2006: 6).

This necessary increase in productivity, like that achieved in manufacturing two centuries ago, is based on the division of labour and on mechanisation. In the localisation industry, division of labour means virtual teams of translators working on a single project, with team members working off-shore to take advantage of lower salaries, or all through different time-zones if what matters is to speed up time-to-market. Mechanisation is achieved through the use of productivity tools: occasionally machine translation (MT), most often TM suites for the translation of running text, and localisation tools for the translation of short strings embedded in programme files. Productivity is achieved through the reuse of already translated text and of its formatting. In fact, it is likely that the savings in formatting reuse are greater than those achieved through text reuse although, surprisingly, no study has been done on this yet.

It is worth noting that the localisation industry does not translate — it *localises.* This involves project managers, graphic designers, software engineers and others working on tasks such as adaptation, quality assurance, desktop publishing adjustments and testing (Esselink, 1998: 258-273), with the translator's role limited to the replacing of natural language strings, a mere, perhaps, 30 percent of the total localisation load. But, yes, this does include the often tough task of respecting the formatting of those natural language strings. It is almost ironic that at the very moment when translation studies was ready to expand the meaning of translation beyond the tight equivalence model that dominated for decades, the localisation industry, the 'market-driven translation theory', moved in the opposite direction, restricting translation to an ('internationalisation-driven) institutionally controlled equivalence (as Pym, 2004: 62–65 explains), thereby giving the translators the added burden of having to go to great lengths to keep the formatting intact.

There are two ways for translators to deal with this formatting issue, and neither is (yet) completely satisfactory. One is overwriting the files, a bad idea if the translator does not have the application with which the original file was created and a working knowledge

of that application. If the file that needs translation is flat, it is not always easy to separate translatable text from code; if it is binary, it won't even be opened without the programme (and, often, the version) that created it. Overwriting is also a bad idea because it does not allow for the semi-automatic reuse of already translated text and, if the translator is lucky, the formatting as well. The other way, which makes more sense for the above reasons, is by using the afore-mentioned TM productivity tools, which translators are forced to do in most localisation projects anyway. The bad news, however, is that, despite claims to the contrary, TM tools do not solve many of these formatting challenges.

A brief look at the users lists for these tools (located at www.yahoogroups.com for most of the best known commercial brands) will show the breadth of the formatting problems translators experience daily when using these tools — and what an advantage it is to be able to count on such quality peer help. Table 1 looks at data from the three lists with the greatest numbers of members and volume of messages for March 2006. Many more queries will have gone to the technical support section in the software developers' web pages or to the language vendors that commissioned the job so the figures are indicative only. The number of messages does not reflect the seriousness of the matters dealt with in them; nor does it reflect on the quality of the particular product. The more 'technical' the job, the more likely it is that there will be more messages dealing with formatting issues. Wordfast, for instance, may have a lower percentage of formatting queries than TRADOS because translators working with Wordfast are likely to do less file-challenging work, not because their software is in any way superior to that of TRADOS.

| List | TW users (TRADOS) | Dejavu-l (DejaVu) | Wordfast (Wordfast) |
|---|---|---|---|
| Formatting-related messages | 107 | 305 | 73 |
| Total number of messages | 448 | 936 | 402 |
| Percentage of formatting related messages | 24% | 32% | 18% |

**Table 1:** Lists and number of formatting-related messages for March 2006

There needs to be better ways of measuring how much time and energy the average translator may use in dealing with formatting glitches. Direct observation of a 'typical' translator's week, now more feasible via usability testing technology, should be attempted to give the research a more controlled, empirical outlook. The author's limited experience as a freelance translator allows him to guess that such kind of research will also confirm that most savings gained through text re-use are offset by the amount of time spent on formatting matters. This article, however, will limit itself to supporting this hypothesis by just looking at the limitations of current technology.

## 3. Current technology promises more than it delivers

Translators receive a job in one of four ways:

1 As files alone.
2 As files plus relevant sections of sentence and term databases.
3 As pre-translated files, with database information inserted in the document, as in pre-translated TRADOS files.
4 As files alone with access to databases hosted in servers.

No system, at present, avoids the problems translators often experience with formatting.

In theory, TM suites and localisation tools separate text from code before translation and then merge translated text with the original code, thus allowing for the reuse of formatting. Then they reuse content, by leveraging data from the databases of translated sentences and terms during the translation process. However, just as these tools don't do automatic translation of the text (TM is not MT!), but just help translators with the repetitive stuff so they are free to concentrate on the more challenging aspects of the text, they don't automatically solve all formatting problems either. The downside here is that dealing with formatting issues and code is a core activity of computer engineers, perhaps desktop publishers or even project managers, not of translators. Therefore, translators are thus less prepared to succeed here.

When language vendors and freelancers encounter the problems related to the reuse of text, they have to deal with formatting too. There are two reasons why they have to deal with formatting: Firstly, because TM

databases are compiled in a proprietary format that does not allow fluid exchange of data with other TM databases — an exchange that is needed as soon as a translator works for a language vendor (or the language vendor for a client) that does not use the same programme. Secondly, because these sentence databases also keep inline formatting (the formatting within the flow of text, as opposed to structured formatting), and a segment with both the right text and format will get a better match than a segment with only the right text.

In fact, exchange of text alone between end clients, language vendors, freelance translators and TM suites is not that difficult. It can always be exported from the database to a spreadsheet programme, then from the spreadsheet to the new database. What is more difficult is working with text that contains both inline formatting and metadata information. When TRADOS became a *de facto* standard in the industry, from the late 1990s onwards, most developers tried to solve the problem by making themselves compatible with TRADOS. Later, when the Translation Memory eXchange (TMX) standard emerged, they all claimed TMX compliance. However, the process of exchanging translation memory data is not always perfect; it was not perfect between TRADOS-compatible software, and it is not even perfect between TMX-certified products at the latest version of the standard, now level 1.4b (Zetzsche, 2003a). In fact, it is developers themselves who simply aim for 'little or no loss of critical data' during the process of exchanging translation memory data (LISA, 2005b).

The problems grow as we move from the reuse of text to the reuse of formatting. At the point of importing a file into whatever translation tool is used, a filter is needed to convert the original file into a format that will be read by the translation editor. Creating these filters and maintaining them throughout the periodic upgrades of the programmes in which the files are composed means a waste of resources for developers — resources that would be better used if devoted to the core function of TM, which is improving the reusability of text.

These problems manifest themselves even further at the point of exporting the file for conversion into its native format, for several reasons:

- conversions are rarely 100 percent accurate
- files may not be well formed due to wrong handling by their creators (for instance, in Microsoft Word, using the enter key to change the line, or the space bar instead of the tab to indent)
- the translator may have pressed the wrong key in the translation editor

Then, we have to account for the possibility of bugs (in the file, in the filter, in the editor), for the difficulties of specific formats (MIF files, resources files), plus possible interferences of hardware / software running in the background.

There is also the issue of text expansion in translation, which will often require post-translation adjustments, particularly in presentation and design-oriented DTP files.

It is relatively easy for the translators and translation project managers to know how these productivity tools should behave in theory. The real test is in the ability of translators and managers to troubleshoot formatting problems as they arise. Allowances for budgeting and time are needed for that, which are likely to eat into most of the savings made through text reuse.

I have not referred here to other issues, such as, for freelancers the maintenance of databases and, for language vendors, the synchronisation of server databases so that they can be effectively used by different translators working at the same time. While time spent in maintenance will also eat into some of the savings from reuse, it is not directly related to formatting.

## 4. Emerging technology: open standards

The problem with formatting is technical and the solution may be technical too. We have seen it emerging through open standards such as the above-mentioned TMX. It is widely accepted that standards benefit everyone — the product developers and businesses that depend on them as well as the actual users — and they have a positive impact on the overall economic cycle. After XML technology was developed, standards were achievable in the area of text reuse, as XML was designed precisely to separate, within the files, content from the container. The Localization Industry Standards Association (LISA) identified this and established in 1997 a specific body to develop text reuse standards. This body is entitled OSCAR (Open Standards for Container/Content Allowing Re-use).

TMX was the first such standard to emerge: version 1.0, for the exchange of text only, was released in December 1997; the latest version, 1.4b, was released

in October 2004 and includes capabilities for exchange of formatting and metadata. All commercial TM suites claim to be compliant with at least version 1.1, while a few certified products, plus some non-certified products, claiming to be compliant with version 1.4b of the standard. There are still the teething problems mentioned above: once again, current software often promises more than it delivers, but the situation is improving.

Term Base eXchange (TBX, version 1.0 released in April 2002) was then developed to cover the terminological exchange needs within the language industry and between tools, not only TM-based oneeds, but also MT-based needs. The Segmentation Rules eXchange (SRX, version 1.0 released in April 2004) followed, once it was realised that up to 30 percent of TMX-exchanged perfect matches could be lost between applications due to differences in segmentation. The last OSCAR standard, still in development, uses the official name of Global Information Management eXchange (GMX), also known as GILT Metrics eXchange. It deals with metrics rather than with text, and consists of three components: GMX Volume for word counting (the only one defined so far), GMX Complexity for the quantification of the complexity of translation tasks, and GMX Quality for the specifications of the quality requirements of translation tasks (LISA, 2005a).

All these OSCAR standards deal with the reuse of text, although GMX only does so indirectly. However, as already discussed, it is in the area of the reuse of formatting that more gains are to be expected from standards. In 2000, a new one standard was developed. It is known as XLIFF (XML Localization Interchange File Format) and comes under the umbrella of OASIS, the Organization for the Advancement of Structured Information Standards. Version 1.1 of XLIFF became an OASIS Committee Specification in the Spring of 2003 (OASIS, 2006).

XLIFF was created for the exchange (OASIS would prefer to call it *interchange* this time) of translatable (or *localisable*) text between different file formats. With XLIFF, content can freely circulate through the localisation cycle with independence of what its native file format was, and independence of the TM suites or localisation tools that will be processed. The XLIFF conversion tool works by separating structural formatting into a skeleton file, then segmenting content and its inline formatting into translation units with its source and its target. These translation units can contain inside 'alternative translation' units, in most cases to hold data leveraged from a TM. Once translated, the XLIFF file merges back with the skeleton to reuse the formatting.

The XLIFF format does much more than simply interfacing with any other file format. It also allows each segment, the minimal discrete unit of translatable text that will then be kept in TM databases for recycling, to carry sophisticated metadata. This metadata can be used to track which version each segment originated from (it is as common for localisation projects to start translation before the final version of the source text has been completed, as it is to update a product, or to generate content from databases instead of static files), and to track which phase of the workflow the segment is going through, including data on tool used, job ID, client, translator/reviser, notes, metrics information, etc. Being an XML standard, it is also extensible and can accommodate future needs (Reynolds & Jewtushenko, 2005).

The XLIFF standard is being developed in line with the OSCAR standards referred to above: segmentation as per SRX rules, TM information so that it can be downloaded from/uploaded to TMX, and word counts based on GMX. Although translation units in the XLIFF format are bilingual only, multilingual projects can be dealt with by bundling together several files in a single document. This is fine, as translation is after all a bilingual activity, and a multilingual file would need to be divided into its bilingual components at some stage anyway.

There is a lot the localisation industry can gain from adopting XLIFF. Complicated projects may have to deal with over thirty different types of files, from EXE and DLL programming files to HTML and XML and their derivatives, to formats generated by content-oriented and design-oriented DTP programs, to the different Microsoft Office applications. Once this standard is adopted, instead of having to build one filter for each file format *plus* filters to handle data between TM suites, software developers will need just one filter for each file format. Indeed, the software that generated the files should produce this filter, thus allowing developers to shift resources to refine the algorithms so that translated text can be reused more thoroughly and easily.

In the current environment, the more that end clients rely on outsourcing localisation to multilingual vendors (MLV) and the less they spend in in-house localisation resources, the more like a 'black box' a whole project looks to them. With current practices end clients pass content and code on to the vendors,

and later receive from the vendors the translated files ready to be imported into their document management system. In an XLIFF environment, clients will have much more control over the whole process, passing only translatable text and keeping the code (which may be sensitive in some cases) in-house. They will gain much more control of their linguistic assets also, just by updating their own TM in the process of converting the XLIFF files to their native format. Just as importantly, they will not risk locking themselves in to a particular vendor or locking in their linguistic assets in to a particular tool.

For language vendors —, particularly those at the top (MLVs) — the success of XLIFF as a standard will mean savings on management, engineering and DTP costs, without having to also lock their linguistic assets in to a particular TM tool. Their current role, which is central in the localisation process, involves dealing with all the formatting complexities the end clients do not want to spend resources on and that the freelance translators do not have the expertise to deal with. It is likely that this role will be transformed into a mere consulting job. SLVs will still retain their important role as language experts, dealing with the linguistic quality assurance of the project.

For freelancers, the success of XLIFF will mean that they will finally be able to concentrate again on text, which is their core activity, rather than on formatting, which is not. It will mean combining the advantages of the pre-digital era, when all they had to worry about was text, with those of the digital era: counting with sophisticated software to re-use translated and repetitive text, allowing them to focus just on the new linguistic challenges arising. No more risks of locking themselves into a particular tool or out from any third party information; no more need to buy several tools for different vendors: any single XLIFF compliant tool will be enough.

## 5. The 100 percent XML scenario

XLIFF may be the next big thing for the localisation industry, as significant as Unicode, which allowed for the easy management of character sets in any language, in any computer and with any (compliant) program. What Unicode did for multilingual writing, XLIFF may do for transporting this written text across languages, localisation agents, software and hardware. The latest development of the past few years of moving client and localisation vendor TM databases from the desktop to the server and triggering the whole localisation process from the client's content management system will be greatly helped by the adoption of XLIFF.

XLIFF, however, is not likely to succeed overnight. At the moment, rather than making all other formats and filters obsolete, it sits there in parallel with them as one more format and one more filter that has to be taken care of by software developers, clients, language vendors and translators, somehow defying the purpose for which it was created. There are also teething problems in the application of the standard, with tools purported to conform to XLIFF producing code that is not easily exchangeable between them (Wunderlich, 2005). Indeed, it may not succeed, just as its OpenTag precedent did not succeed. Not enough end clients and leading software developers may feel the need to invest the resources needed to make it run. Some MLVs (gatekeepers as they are sometimes known) may resist it as it makes almost obsolete what is now a big chunk of their core activity. Like all standards, XLIFF has been developed by big players — with Novell, Oracle, Sun, and Berlitz involved first, then joined by Lotus/IBM, Moravia IT, RWS Group, and Lionbridge — but that does not guarantee its success.

On the other hand, XLIFF seems to be making inroads into the industry. Leading commercial localisation tools (Catalyst, PASSOLO, WinRC) and TM suites (SDLX/TRADOS, Heartsome) have adopted it. There is interest in the open source community in the use of this standard (Frimannsson & Hogan, 2005), with KBabel and Language Tools also offering free XLIFF-compliant tools. In some cases (SDLX for instance) the XLIFF format will interface with the native file format via its own proprietary ITD format. Heartsome, on the contrary, works directly on XLIFF, TMX and TBX standards without using any proprietary standards. Innovators and early adopters are already embracing XLIFF, although we are still at the first stages of the S-curve. For clients and language vendors, there is no longer any comparative advantage in adopting TM as most are already using it so rather than TM being advantage it is a necessity. However, there may be a comparative advantage in adopting XLIFF, and server TM and document management software now, before the majority does (Project-Open, 2005).

Indeed, it is easy to imagine a 100 percent XML scenario in which a more developed XLIFF specification would be able to carry out the management of information of the global enterprise seamlessly — from the authoring of text to its localisation, publishing and archiving, with processes triggered and pushed through the corresponding

workflow (semi) automatically by content management software, all overseen by the project manager. Technical writers will create content on structured language and, with the help of authoring tools, through the single sourcing cycle, allowing for text chunks to be reused in other documents and to be outputted in different formats: HTML, PDF, Help, etc. (Rockley, 2002). Then, translators will move it through the localisation cycle while reusing previously translated sentences and terms. Both, technical writers and translators — language specialists in their own right— will deal only with text and, when relevant, its recycling, leaving formatting to DTP and engineer specialists who will handle it in a totally independent way.

In this scenario, successful software developers could actually afford the resources to enhance text reuse algorithms that incorporate linguistic knowledge (inflections, synonyms…) and perhaps create a new kind of language-specific TM which is more efficient for the particular language combination. Doing this would blur the distance between MT and software development, but still leave translators in charge. The process could then rightly be considered as machine assisted human translation rather than human assisted machine translation to use Hutchins' (1992) parlance. For content creation, translation, and translation management some developers may find it useful to pursue Zydron's (2003) 'text memory' xml:tm idea. Others may be interested in advancing diagnostic tools to determine whether a document should be translated by MT, by TM or without them, as the TransRouter project (Cleary & Schaler, 2000) was aiming at.

With this 100 percent XML scenario, just as in pre-digital times, translators will need to focus only on text, which is complex enough, without being distracted by the complications of formatting. After all, there will be enough challenges for freelancers in coping with the demands of translating following the imminent introduction of Web 2.0 — the Semantic Web in which machines, rather than merely displaying data as they do now, will be better able to 'understand' it as well (Berners-Lee, Hendler, & Lassila, 2001). Even if XLIFF succeeds, the digital world will still stir the translation profession for years.

**References**

Arevadillo Doval, J. J. (2005). The EN-15038 European Quality Standard for Translation Services: What's Behind It? *The Globalisation Insider.* Retrieved 1 April 2006, from http://www.lisa.org/globalizationinsider/2005/04/the_en15038_eur.html

Austermuhl, F. (2001). *Electronic Tools for Translators*. Manchester: St Jerome.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. *Scientific American,*(5).

Bowker, L. (2002). *Computer-Aided Translation Technology: A Practical Introduction*. Ottawa: University of Ottawa Press.

Chesterman, A., & Wagner, E. (2002). *Can Theory Help Translators*. Manchester, UK & Northampton, MA: St Jerome Publishing.

Cleary, R., & Schaler, R. (2000). R&D Opens Doors to Translation Portals. *Localisation Ireland, 4*(1), 9. DePalma, D. A., & Beninatto, R. (2006). *Ranking of Top 20 Translation Companies for 2005*: Common Sense Advisory.

Esselink, B. (1998). *A Practical Guide to Software Localisation.* Amsterdam & Philadelphia: John Benjamins.

Fenstermacher, A. (2006, January/February). Authors, localizers and language barriers. *Multilingual, 77,* 82.

Frimannsson, A., & Hogan, J. M. (2005). Adopting Standards-based XML File Formats in Open Source Localisation. *Localisation Focus – The International Journal for Localisation, 4*(4), 9–23.

Hutchins, J., & Somers, H. (1992). *An Introduction to Machine Translation*. London: Academic Press.

LETRAC. (1999). LETRAC Curriculum Modules. Retrieved 1 April 2006, from http://www.iai.uni-sb.de/docs/D3.pdf

LISA. (2005a). OSCAR (Open Standards for Container / Content Allowing re-use). Retrieved 1 April, 2006, from http://www.lisa.org/sigs/oscar/

LISA. (2005b). TMX - Translation Memory eXchange. Retrieved 1 April, 2006, from http://www.lisa.org/tmx/

OASIS. (2006). OASIS XML Localisation Interchange File Format (XLIFF) TC. Retrieved 1 April, 2006, from http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=xliff

Project-Open. (2005). Technology Adoption and Competitiveness in the Translation Industry. Retrieved 1 April, 2006, from http://www.project-open.com/whitepapers/index.html#Liese2005

Pym, A. (2004). *The Moving Text. Localization, Translation, and Distribution*. Amsterdam and Philadelphia: Benjamins.

Reynolds, P., & Jewtushenko, T. (2005). What Is XLIFF and Why Should I Use It? *XML Journal*, 4.
Rockley, A. (2002). *Managing Enterprise Content: A Unified Content Strategy*. Indianapolis: New Riders Publishing.

Somers, H. (Ed.). (2003). *Computers and Translation: A translator's guide* (Vol. 35). Amsterdam/Philadelphia: John Benjamins Publishing.

Wunderlich, M. (2005). Options for Editing an XLIFF File. *Multilingual Computing and Technology, 76*, 51–58.

Zetzsche, J. (2003a, March 2003). TMX Implementation in Major Translation Tools. *Multilingual Computing and Technology, 14, 2,* 23–27.

Zetzsche, J. (2003b). *A Translator's Tool Box for the 21st Century. A Computer Primer for Translators* (2nd Version 2.4.1, December ed.): International Writers' Group.

Zydron, A. (2003, August 26, 2003). xml:tm Using XML technology to reduce the cost of authoring and translation. *The LISA Newsletter, XII*.

# Beavers, Maple Leaves and Maple Trees

**Julie McDonough**
**University of Ottawa**
**Canada**
**julielaura.mcdonough@uottawa.ca**

Because a national symbol appeals to the sense of collective identity shared by the members of a nation, its use in localised websites by companies from outside the nation merits reflection. In this paper, a case study of thirty of the largest American corporations is used to explore how common it is for national symbols to be incorporated into websites localised for Canadian users. The results are then compared to the use of national symbols on the websites of thirty of the largest Canadian corporations to determine whether national symbols are adopted more frequently by domestic or international companies. The paper ends with some reflections on the inclusion of national symbols within a localised website and the ambiguity of their meaning.

**Keywords:** *localisation, collective identity, symbolism, national identity, website translation, adaptation, national symbols*

## Introduction[1]

Given that the purpose of localisation is to ensure that localised products are adapted to the conventions of a given locale (Lommel 2003, p.5), a successfully localised website should not immediately strike targeted users as being different from a site designed by a domestic company. As part of the process of adapting the site for a new locale, especially when targeting an entire country, localisers may decide to incorporate images of locale-specific national symbols such as flags or monuments. Singh and Pereira (2005), for instance, recommend including symbols and "pictures of national identity" such as those that emphasise architectural achievements or national pride on websites designed for collectivist locales (2005, p.83). They stress that domestic companies will be incorporating such symbols in their own websites and advertisements, implying that international companies will be at a disadvantage if they do not follow suit.

Yet this assumption may merit further reflection. This paper will raise questions about the use of national symbols on websites and explore the issues surrounding the inclusion of national symbols within localised sites. It will do this by first exploring what national symbols are supposed to represent and then studying examples of Canadian symbols used in localised and domestic websites.

## 1. National symbols

National symbols help form and maintain national identity, mark a nation's collective memory, preserve its shared past and represent the power of a state to define a nation (Geisler 2005, pp.xv–xvi). In a way, national symbols act much like the logo of a corporation, as they are a means by which the State can depict an image of itself to members and outsiders alike. Much like a corporate logo, national symbols represent a nation's key values and are chosen because they have special significance for the nation and its members. As Smith (1991, p.77) notes, national symbols, customs and ceremonies make the concepts of a nation visible for all members and appeal to their emotions.

However, what exactly constitutes a national symbol is not unanimously agreed upon. Smith (1991, ibid.), for instance, groups symbols, customs and ceremonies together and considers flags, anthems, parades, coins, capital cities, folk costumes, folklore museums, war monuments, passports and borders to be "obvious" examples. Cerulo (1995, p.13) adds mottoes and shrines to this main list, and Smith later expands his initial examples with a series of "hidden" ones, including popular heroes or heroines, fairy tales, educational practices and military codes. He asserts that these symbols, customs and ceremonies are the ways of acting shared by a "community of historical culture" (1991, p.77). Geisler, on the other hand, argues that Smith's typology may be too broad. He suggests that a narrow typology of important national symbols would minimally include the flag, anthem, national holidays, currency, capital and major national monuments, with the flag being the most important and the others ranked somewhere below it (2005, pp.xxi-xxii).

[1]The author would like to thank Clara Foz for her feedback on an earlier version of this paper.

These symbols, Geisler asserts, serve to support and reinforce a nation's identity, both within its borders and to the outside world. Each time such a symbol is "actualized" — whenever an anthem is sung or a flag is raised, for instance — it reminds members of the nation that they share a common past and are bound by a collective identity. In fact, only through constant repetition of a symbol in the media, political speeches, public ceremonies, etc. do members of a nation become attached to it (Geisler 2005, pp. xix, xxvii). Similarly, David Bell (2001, p.95, following Anderson 1983) suggests that nations are imagined communities and that symbols are essential for this community to become a nation, since members can interact with others only through shared 'things' such as an anthem or flag and a set of customs and rituals.[2]

Because national symbols depict a nation's history, values and identity, they send an ambiguous message to users when they are incorporated into a localised website. On one hand, the symbols act as a sort of logo that identifies the site as the Canadian version. A Canadian flag beside a "change location" link may simply serve to distinguish this site from one that has, say, a French, Japanese, or Chilean flag, alerting users that they are indeed browsing the site designed for their locale. The ambiguity results from the fact that the 'imagined community' described by Bell and Anderson is also projected by localised websites incorporating local symbols. Just as a logo stamped on a product's packaging signals to consumers that the product has been made by and comes from a given company, so do national symbols signal that a website has been made in and is part of a given nation. Users may therefore interpret a national symbol to mean that both the company and the user belong to the same imagined community, share the same collective identity, and are bound by the same common past.

## 2. Use of Canadian symbols

The Government of Canada officially recognises three national symbols in addition to the national flag, colours, seal, and anthem: the beaver, maple tree and maple leaf (Government of Canada, Canadian Heritage 2004). All of these symbols are infused with special historical significance. The beaver, which became "a symbol of the sovereignty of Canada" when the National Symbol of Canada Act was passed in 1975, is a reminder of the importance of the fur trade to the early Canadian economy in the 17th, 18th and 19th centuries, when thousands of Canadian beaver pelts were shipped to Europe annually for use in fur hats. It appears on the Canadian five-cent coin and was featured on the first Canadian stamp. The maple tree, officially recognised as Canada's "arboreal emblem" in 1996, has "played a meaningful role in the historical development of Canada and continue[s] to be of commercial, environmental and aesthetic importance to all Canadians" (ibid). At least one of the ten species of maple native to Canada grows in every province, and the sap is used to make maple syrup, of which Canada is the leading global producer.[3] Finally, the maple leaf (Figure 1), incorporated into the Canadian and Ontario flags in 1965, appears on the one-cent coin and is featured in The Maple Leaf Forever, a song composed for Canada's confederation in 1867 and an unofficial English-Canadian anthem for several decades.

**Figure 1:** The 11-point maple leaf, an official symbol of Canada.

Though these symbols may be accorded official State-recognised status, several others could be considered to have semi-official status, even if one follows only the narrow typology of symbols offered by Geisler (2005). One could reasonably include the moose, found on the twenty-five cent piece; the loon, depicted on the one-dollar coin; the polar bear, which appears on the two-dollar coin; and the Bluenose, a fishing schooner built in the 1920s that was renowned for winning several international races during the 1920s[4], represented Canada at the 1933 Chicago World Fair, was sent to England on behalf of Canada in 1935 for the Silver Jubilee of King George and Queen Mary, and is featured on the ten-cent coin (Province of Nova Scotia, Department of Tourism, Culture & Heritage

[2] See also Hall (1996) for further discussion of imagined communities and cultural representation
[3] Canada is responsible for approximately 85 percent of world maple syrup production. See fact sheet available at: http://ats.agr.gc.ca/supply/3310_e.htm [accessed 12 April 2006].
[4] It was in fact dubbed the "Queen of the North Atlantic fishing fleet".

2004). And since Geisler also includes major national monuments, one might add to this semi-official list the Canadian Parliament buildings or the War Memorial in Ottawa, where Remembrance Day ceremonies are held each year.

Other regions of Canada have their own provincial or 'national'[5] symbols as well. All of the country's ten provinces and three territories have an official coat of arms, flag and flower and many regional groups, such as Franco-Ontarians, also have officially recognised flags or emblems. While such symbols could also be used by localisers to target a website to a specific group of Canadians, this paper will focus only on national symbols representative of Canada as a whole rather than a particular region. A future study will encompass a wider range of symbols, as their use on a website will help indicate which particular segments of the Canadian population a company may be trying to target. For the purposes of this case study, both the official and non-official symbols mentioned in the two preceding paragraphs were considered to be national symbols of Canada.

### 2.1 Case study: Methodology

To study the use of Canadian national symbols on localised websites and those of Canadian companies, thirty of the largest American corporations and thirty of the largest Canadian were compared. A Canadian company has been defined as one that has its headquarters in Canada and is not a subsidiary of an international company. Sears Canada, for instance, would be considered Canadian even though Sears Roebuck owns more than 50% of its shares[6].

The American companies were selected based on the Fortune 500 list published by Fortune magazine on 18 April, 2005, while Canadian companies were selected based on the 2005 Top 1000 Companies rankings compiled by Report on Business Magazine, which is published by The Globe and Mail, a major Canadian daily newspaper.

The Fortune and Globe and Mail rankings were chosen for two reasons. First, given the fact that localisation involves a considerable investment of financial and human resources, larger corporations are more likely than smaller companies to have international operations and localised websites for foreign markets. In addition, the corporations that head the list fall under various industries, making the sample more representative of large Canadian and American corporations in general rather than of those in a particular sector. Though energy companies do figure prominently in both lists[7], Fortune's top fifty also includes department stores such as Wal-Mart and Costco, speciality stores such as Home Depot, and manufacturers such as Procter & Gamble, Ford, and Dell, while the Globe and Mail top fifty includes banks such as CIBC, RBC, BMO and TD, grocery retailers such as Loblaws, telephone utilities such as Bell Canada, and manufacturers such as Magna International.

The American companies were selected from the top fifty-three of the Fortune 500, beginning with Wal-Mart (#1) and ending with Merrill Lynch (#53). In order of ranking, the website of each corporation was checked, and if a Canadian version of the site was available, the company was included in the study. Twenty-three of the top fifty-three companies had to be excluded as no Canadian version of their website was available[8]. In each case, the next-ranked company was chosen so that a total of thirty could be included in the case study. Websites were considered localised for Canada when the US parent company had a global gateway from which a 'Canada' or 'Canadian' site could be accessed or when a link to the Canadian version was posted on the American website. When English- and French-Canadian sites were available, both versions were consulted; otherwise, the English-Canadian site was considered to be the localised version[9]. In total, thirty-seven websites representing the thirty companies and their subsidiaries were consulted.

As a point of comparison, thirty Canadian companies were selected from among the first thirty-nine on the R.O.B. Top 1000 Companies list. For the purposes of this study, when both a holding/parent corporation and its subsidiaries were listed, they were not counted as separate companies, though the websites of both the

[5] Though other provinces or territories usually use the term 'national' to refer to the federal government and Canada as a whole, Quebec often uses the term to refer to Quebec institutions and symbols. Thus, Quebec's provincial legislature is referred to as the *Assemblée nationale* or National Assembly, the Quebec government's highest award of distinction is the *Ordre national du Québec,* the region surrounding Quebec's provincial capital is referred to as the *capitale-nationale* and Saint-Jean-Baptiste Day, an official holiday only in Quebec, is referred to as the *Fête nationale.*[3] Canada is responsible for approximately 85 percent of world maple syrup production. See fact sheet available at: http://ats.agr.gc.ca/supply/3310_e.htm.

[6] The companies listed in the ROB report are all publicly traded on the Toronto Stock Exchange. The problems that arise from this definition will be explored in the next section.

[7] Of the top thirty-nine companies on the Globe and Mail list, for instance, ten are classified as oil- and/or gas-related (oil and gas producers, integrated oil, gas pipelines, etc.).

[8] e.g. Bank of America (#18), Target (#27), Morgan Stanley (#36) and Metlife (#37).

[9] A small percentage of both the localised and domestic sites were available only in English:18 of the 52 domestic and 6 of the 37 localised.

parent and the subsidiary were consulted. For instance, Power Corp (#26) holds Power Financial Corp (#13), which in turn holds Great West Lifeco (#20). All three corporations were counted as only one of the thirty in this case study, though the website of each was examined. In addition, whenever an additional Canadian version of a website was separate from the main corporate site, both were consulted (e.g. www.loblaw.com, the corporate website for Loblaw Companies Limited, was consulted, as were the websites of its retail locations and brands, including Loblaws, No Frills, and President's Choice). In total, fifty-two websites were consulted, representing thirty companies and their subsidiaries. These websites were considered 'domestic' in contrast to the 'localised' sites of the American corporations.


**Figure 2:** Maple leaf incorporated into a logo

A company was considered to be using a national symbol on its website when the symbol was part of the company logo (Figure 2), appeared as part of the background image, was beside a link to a page within the website, or was included in an image on one of the pages within the site (Figure 3). Companies were not considered to be using a national symbol when it appeared on the website because it was obviously part of a logo or image of an outside source. For instance, the homepage of RBC Financial Group, a Toronto-based financial corporation, has a small image in the lower-right-hand corner indicating that RBC was named 'Canada's Most Respected Corporation' for 2005[10]. Above this statement is a copy of the design etched onto the base of the trophy. Although this design includes a red maple leaf, RBC was not considered to be using a Canadian symbol because the image came from the sponsor of the survey, not RBC. The image was therefore not designed by or on behalf of RBC and served only to link to a news article about the award and to the survey website: http://www.mostrespected.ca/.

## 2.2 Case Study: Findings

Of the thirty American companies with localised websites studied for this paper, a total of twelve incorporated Canadian symbols. Maple leaves were used by six of these companies, while the Canadian


**Figure 3:** Example of a Canadian symbol included within an image on a website

flag was used by the other six. American International Group, which owns AIG Life, used a Canadian flag on the AIG website and a maple leaf on that of AIG Life. Only one site, General Motors Canada, used both a maple leaf — as part of its logo — and a flag, while General Electric included both maple leaves and a photograph of the CN Tower, arguably a Canadian national monument, as it is billed as "Canada's wonder of the world" on the CN Tower website (www.cntower.ca). No other national symbols (e.g. beaver, Parliament) appeared to be used by any of the companies. Table 1 summarises the use of national symbols on these websites.

The websites of the thirty Canadian companies and their subsidiaries also included Canadian symbols, to almost the same extent: thirteen of the fifty-two sites — representing eleven of the thirty companies — had images of maple leaves or trees, the Canadian flag, a beaver[11] or the CN Tower on their sites. Many of the websites in this group were the corporate sites intended both for Canadians and international users and hence would not necessarily focus on the company's 'Canadianness'; however, some of the .com sites included national symbols, while many of the .ca sites did not. For instance, none of the Loblaw subsidiaries or brands — including No Frills, Fortinos, Maxi, Zehrs Markets and Independent — used Canadian symbols, though each of these latter sites has been localised for users within the province(s) where that chain of grocery stores is located. This shows that a site does not have to be targeted to just Canadians for a company to highlight its Canadian roots. Table 2 summarises the use of national symbols on these websites.

[10] Manulife Financial has this same image on its homepage..

[11] While the current Bell Canada site does not have any Canadian symbols, the beta version, which was available for a short time in early 2006, included the only image of a beaver found on any of the sites in this case study. These beavers, named Frank and Gordon, are part of a larger advertising campaign. They can still be found at a Bell microsite: http://www.frankandgordon.ca/.

| Fortune Ranking | Company | Canadian website | Symbol(s) | Location of symbol(s) |
|---|---|---|---|---|
| 2 | Exxon Mobil | http://www.exxonmobil.com/Canada-English/HR/HR_Can_Homepage.asp | Canadian flag | HR page (home page of Exxon Mobil Canada. The site has been partially localised.) |
| 3 | General Motors | www.gmcanada.com | 1. Maple leaf | 1. In GM Canada logo |
| | | | 2. Canadian flag | 2. In a photo GM dealership on the Site Map/About Us pages |
| 5 | General Electric | www.ge.com/ca | 1. Red maple leaves | 1. Homepage |
| | | | 2. Green maple leaf | 2. Homepage |
| | | | 3. Photo of CN Tower | 3. Our Company page |
| 9 | American International Group | http://www.aig.com/gateway/home/1-113-Canada_index.htm | Canadian flag | Beside 'change location' link on navigation bar |
| * | | http://www.aiglife.ca/ | Red maple leaf | On homepage, beside 'AIG Life of Canada' |
| 19 | State Farm Insurance | www.statefarm.ca | Red maple leaf | On homepage, beside 'statefarm.ca' |
| 24 | Pfizer | www.pfizer.ca | Red maple leaf | On homepage, beside 'healthcare in Canada' heading |
| 26 | Procter & Gamble | http://www.pg.com/en_CA/index.jhtml | 1. Red maple leaf | 1. Start-up splash page |
| | | | 2. Red maple leaf | 2. On homepage, beside link to 'P&G global operations' |
| 28 | Dell | http://www.dell.ca/ | Canadian flag | Beside Dell Canada logo on navigation bar |
| 30 | Johnson & Johnson | www.jnjcanada.com | Canadian flag | Homepage |
| 32 | Time Warner (AOL, Time) | www.aol.ca | Red maple leaf | Beside search bar. |
| | | | | Note: maple leaf not on Quebec site (www.aol.qc.ca) |
| 52 | Wells Fargo | http://financial.wellsfargo.com/canada/en/index.html | Canadian flag | Beside Search text box |
| 53 | Merrill Lynch | http://gmi.ml.com/canada/ | 1. Canadian flag | 1. In an image under navigation bar behind the word 'Canada' |
| | | | 2. Canadian flag | 2. In an image on homepage |
| | | | 3. Grey Maple leaf | 3. Watermark background image on homepage |

**Table 1:** National symbols used in websites localised for Canada

* Each of these companies was considered a subsidiary of the Fortune - or Globe and Mail - ranked company just above it (e.g. Bell Canada and Telesat are subsidiaries of BCE Inc.). With the exception of Bell Canada,these companies did not appear in the Fortune 500 or Globe and Mail top 1000.

| G&M Ranking | Company | Website | Symbol(s) | Location of symbol(s) | Head office |
|---|---|---|---|---|---|
| 7 | Canadian Imperial Bank of Commerce | http://www.cibc.com/ca | Photo of CN Tower | CIBC world markets page | Toronto, Ontario |
| 10 | Petro-Canada | www.petro-canada.ca | White maple leaf | Used in logo | Calgary, Alberta |
| 14 | BCE Inc. | http://www.bce.ca/ | – | | Montreal, Quebec |
| * | Bell Canada (#15) | www.bell.ca | – | | Montreal, Quebec |
| * | | http://www.telesat.ca/ | Red maple leaf | Used in logo | |
| 16 | Cdn. Natural Resources | www.cnrl.com | White maple leaf | Used in logo | Calgary, Alberta |
| 18 | Canadian National Railway Co. | www.cn.ca | Canadian flag | On the homepage, in an image of a small, red CN train filled with people. Both US and Canadian flags are flying on the train. | Montreal, Quebec |
| 19 | Shell Canada | www.shell.ca | 1. Orange maple leaf | 1. Photo to mark the link to the 'code of ethics' page under About Shell – How we work | Calgary, Alberta |
| | | | 2. Canadian flag | 2. In a photograph of Shell's Peace River Complex (Shell for Businesses – Natural Gas & Co-products – Asphalt) | |
| | | | 3. Red maple leaves | 3. Photo to mark the link to the 'Shell Canada's Core Values' page under Jobs & Careers – Working for Shell Canada | |
| * | | http://www.sunoco.ca/ | Canadian flag | Within a Sunoco poster advertising the 'Ron Fellows Karting Championship' on the Community page. The poster is part of an image of race cars speeding around a corner | |
| 24 | Husky Energy | www.huskyenergy.ca | Maple tree branch | Photo on the About Husky – Health Safety & Environment page | Calgary, Alberta |
| 31 | Nexen Inc. | http://www.nexeninc.com | Red maple leaf | Used in logo | Calgary, Alberta |
| 34 | Talisman Energy | www.talisman-energy.com/ | Photo of a maple tree | On the About Us page | Calgary, Alberta |
| 35 | Enbridge Inc. | www.enbridge.com | Photo of Enbridge van with CN tower in far background | On the Library page | Calgary, Alberta |
| 40 | Ipsco Inc. | www.ipsco.com | Canada goose | Used in logo | Regina, Saskatchewan |

**Table 2:** National symbols used in the websites of Canadian companies

* Each of these companies was considered a subsidiary of the Fortune - or Globe and Mail - ranked company just above it (e.g. Bell Canada and Telesat are subsidiaries of BCE Inc.). With the exception of Bell Canada,these companies did not appear in the Fortune 500 or Globe and Mail top 1000.

As Tables 1 and 2 illustrate, Canadian symbols are used on both domestic and localised websites. In both sample groups, the maple leaf and Canadian flag were favoured over other official or semi-official symbols, and though these emblems were most commonly located on the homepages of localised sites and in the logos of domestic sites, they were also found on various other pages.

The results raise intriguing questions, among which are what the national symbols are intended to represent and what their function is supposed to be. On the localised sites, one can reasonably assume that a Canadian flag — as used on the Canadian version of the AIG website, for instance — is intended to help users distinguish one locale from another and signal that the parent company is making an effort to be part of the Canadian community. This assumption is supported by the fact that of the thirty-seven localised websites in this study — including those that did not incorporate national symbols — only two, those of General Motors Canada and AIG Assurance[12], did not appear to have a link to their parent company. Thus, the fact that these localised websites are part of a larger, global operation is not actively concealed from users in the targeted locale, regardless of whether or not national symbols are used.

Yet the function of national symbols is not necessarily the same in the domestic websites. Because a Canadian-owned company is actually part of the Canadian 'imagined community', the national symbols on its website signal not only that the site is intended for the English- and/or French-Canadian locales, but also that both the company and Canadian users share the same imagined identity. The Canadian symbols create a bond (see Cerulo 1995, p.16) between the company and the user, appealing to the latter's sense of collective identity, belonging and patriotic desire to support local businesses. The symbol may or may not achieve this effect, but it certainly performs this function.

And here lies the issue upon which further reflection is merited. Unless a user actively searches though a website to determine whether the company is in fact Canadian, how is he or she supposed to know what the national symbol is intended to represent? When an image of a maple leaf, Canadian flag, or Canadian monument is found on a home or start-up splash page of a .ca website, a user's first inclination would be to identify the company as Canadian, whether or not this is actually the case.

The issue is made more complicated by the fact that in some ways, a subsidiary of an American company is still a part of Canada, though not technically owned by Canadians[13]. Both GE and General Motors Canada, for instance, have long histories in Canada and employ thousands of Canadians. GE's first manufacturing facility in Canada was opened in 1892, while General Motors Canada was established when GM bought the family-run and Canadian-owned McLaughlin Motor Car Company in 1918[14].

In other cases, a Canadian company, though not a subsidiary of a larger, international operation, may not be entirely Canadian-owned. As mentioned earlier, Sears Canada is not a subsidiary of Sears Roebuck[15], since Sears Canada was actually formed as a 50-50 partnership between The Robert Simpson Company, a Canadian retailer, and Sears Roebuck in 1953. However, Sears Roebuck has since increased its ownership of Sears Canada: in 1984 it held 62.6% of the company, but by 1996 it held a smaller majority of 55% of shares[16]. As ownership changes hands over time, does a Canadian company become more or less Canadian? And if so, should its 'right' to use Canadian symbols be revoked? Corporate ownership is often difficult to precisely determine, which only adds to the ambiguity surrounding what national symbols really represent on commercial websites.

Even the legislation related to the use of Canadian symbols does not completely elucidate the issue. Several symbols are protected by Canadian law. The national flag and coat of arms, for instance, are protected by The Trade Marks Act, which forbids commercial use of these symbols without permission from the federal government's Department of Canadian Heritage.[17] The maple leaf itself is protected by both an international treaty (Paris Convention for the Protection of Industrial Property) and Canadian legislation (Order in Council P.C. 1965–1623), though the exact symbol referred to is the 11-point maple leaf (Figure 1) that appears on the Canadian flag. Likely

[12] GM Canada does, however, outline its history in Canada and its links to General motors. See: http://www.gmcanada.com/inm/gmcanada/english/about/OverviewHist/hist_gm_canada.html [accessed 17 April 2006], while AIG Assurance details its relationship to AIG at http://www.assuranceaig.ca/aboutus.asp [accessed 17 April 2006].
[13] Thanks to the referee of a previous paper for this point.
[14] Moreover, Sam McLaughlin remained president of GM Canada until 1945 and Chairman of the board until his death in 1972. Full history available at http://www.gmcanada.com/inm/gmcanada/english/about/OverviewHist/hist_gm_canada.html [accessed 17 April 2006].
[15] #45 on the Fortune 500 list.
[16] See the Sears History feature available at http://www.sears.ca/e/corporate/about_home.html# [accessed 17 April 2006].
[17] See the Government of Canada, Canadian Heritage page at http://www.pch.gc.ca/progs/cpsc-ccsp/sc-cs/commuse_e.cfm for further details.

because of these regulations, the maple leaves included on sites such as Procter & Gamble or Merrill Lynch are not exactly the same as the 11-point trademarked symbol protected by the Canadian government. In this way, the symbol itself can still be used in an effort to create a bond between the corporation and website users without actually infringing on trademark laws.

Yet, even if the maple leaves incorporated into a localised or domestic website are not identical to the 11-point leaf officially recognised and protected by the Canadian government, they still function in much the same way as their official counterpart. A maple leaf — and other Canadian symbols — will appeal to a user's sense of collective identity, regardless of whether or not it has eleven points and is identical to the one on the Canadian flag. And because any national symbol will operate on more than one level, no company can be sure that it will be received and interpreted as intended.

Consider, for instance, the fact that in Canada national symbols do not evoke the same reaction from all Canadians. Supporters of Quebec independence or sovereignty often view the Canadian flag and maple leaf negatively. The Parti québécois, a secessionist provincial political party, for instance, once refused federal funding for renovations to the Quebec City zoo and aquarium because the grant was tied to the condition that bilingual signs be posted and the Canadian flag fly over both buildings for forty years. This offer was decried as "une tentative de relativiser notre statut national" ["an attempt to dilute our national status"[18]], and the Quebec government instead funded the entire 38 million dollar project itself (Lessard 2001, p.A1; Séguin 2001, p.A01). Yet only one of the sites in this study seemed aware that the maple leaf could potentially have a negative, rather than a positive, effect on a user's reception of the local site: while the AOL English-Canadian website included a red maple leaf, this symbol was not found on the AOL French-language Quebec site, though no official Quebec symbols such as the flag or fleur-de-lys were used instead[19]. AOL thus differentiates between English Canadians, who are expected to be receptive to the maple leaf, and French Canadians, who may not be.

National symbols, then, risk not only being misinterpreted by users, but also, in some cases, acting contrary to the corporation's intention: instead of creating a bond between users and the company, red maple leaves and Canadian flags may actually alienate certain segments of the intended audience, who do not consider the national flag and other official emblems representative of their national identity.

Yet, using the official symbols of a particular group of Canadians (e.g. fleur-de-lys for Quebecers, especially the French speakers) to better reach a group that feels little attachment to the national symbol would simply create more problems, since an additional site would have to be created to target this locale. Instead of offering English- and French-Canadian sites, a company would ideally have to create one for English Canada, one for French Canada (since French speakers live throughout the country) and yet another for Quebec, available in at least two languages, as both English and French speakers reside within the province. Localisers would be creating largely unnecessary segmentation and additional websites simply to include various official symbols that may not even be received as intended. And the smaller the group targeted by the localised site, the less likely the company is to see a significant return on its investment.

As the results of this case study indicate, Canadian and American companies do not uniformly use national symbols on their (localised) websites. Since approximately half of both the American and Canadian companies included some officially recognised symbol, it is unclear what the symbols are supposed to represent. At best, they are used haphazardly by companies and are included or removed when sites are redesigned. No national symbols were found on the current Wal-Mart Canada site, for instance, though in 2005 a red maple leaf appeared on the home page beside the "Welcome to Wal-Mart Canada Corp." heading. Canadian website users are therefore receiving unclear messages about a company's status in Canada. They may not realise that a Canadian company without a national symbol on its site is in fact Canadian, and they may mistakenly believe an American company is Canadian owned or has its headquarters in Canada simply because it uses maple leaves, maple trees or the Canadian flag somewhere on its website.

## Conclusion

As discussed, national symbols in localisation have a dual nature. Superficially, they are accessories used to

[18] English version from the Globe and Mail (Séguin 2001, p.A01)

[19] Appropriately so, since French speakers live in several provinces and not just Quebec and some French-speaking users of the AOL French-Canadian page would therefore not identify with the fleur-de-lys or Quebec flag.Table 1: National symbols used in websites localised for Canada

designate the locale for which a website has been designed. But because they also have a more figurative function — that of reaffirming collective identity — they may be interpreted by some users as an indication that both the company and the user are part of the same imaginary nation represented by the symbol and thus share the same historic roots and core values.

The goal of localisation, notes Yunker (2003, p.18), is not to "trick" users into thinking a company is local, but rather to let them know that the company understands the needs and wants of users in a given locale. Are national symbols necessary for transmitting this latter message? Not necessarily. In fact, I would argue that precisely because localisation is not supposed to deceive users, a corporation should carefully consider how national symbols might be interpreted before deciding whether or not to include them in a localised website. Localisers would also be wise to consider the political implications of incorporating federal symbols into websites when such symbols risk alienating or, at the very least annoying, users in the locale.

National symbols are not the only way of appealing to a locale in which collective values are very strong. Focus can still be placed on the company's place in and contributions to the area by highlighting its involvement in the local community, its donations to local charities, the number of jobs it has created within the region, etc. In this way, the chance for users to misinterpret a company's intentions or origins would be reduced and fewer users would likely be antagonised, while the company's contributions to the locale would not be overlooked.

Experimental research into user reception and interpretation of symbols in websites would complement this study and help provide more definite conclusions about whether these emblems are being interpreted as localisers intended.

**References:**

Anderson, B. (1983) *Imagined Communities: Reflections on the Origin and Spread of Nationalism*, London & New York: Verso.

Bell, D. (2001) *An Introduction to Cybercultures*, New York: Routledge.

Cerulo, K. A. (1995) *Identity Designs: The Sights and Sounds of a Nation*, New Brunswick, New Jersey: Rutgers University Press.

Geisler, M. E. (2005) 'What Are National Symbols and What do They Do to Us?' in Geisler, M. E., ed., *National Symbols, Fractured Identities: Contesting the National Narrative,* Middlebury, Vermont: Middlebury College Press, xiii-xlii

Globe and Mail (2005) *Top 1000 Companies* [online], available: http://www.globeinvestor.com/series/top1000/tables/companies/2005/ [accessed 7 April 2006].

Government of Canada, Canadian Heritage (2004) *The Symbols of Canada* [online], available: http://www.pch.gc.ca/progs/cpsc-ccsp/sc-cs/index_e.cfm [accessed 8 April 2006].

Hall, S. (1996) 'The Question of Cultural Identity' in Hall, S., Held, D., Hubert, D. and Thompson, K., eds., *Modernity: An Introduction to Modern Societies*, Malden, Massachusetts: Blackwell Publishers, 595-634.

'Largest U.S. Corporations' (2005), *Fortune*, 151(8), F1-F20.

Lessard, D. (2001) 'Le drapeau canadien, « un bout de chiffon rouge »! lance Landry', *La Presse,* 24 jan, A1.

Lommel, A. (2003) *The Localization Industry Primer*, 2nd ed., Féchy, Switzerland: Localization Industry Standards Association [online], available at: http://www.lisa.org/products/primer.html [accessed 8 August 2005].

Province of Nova Scotia, Department of Tourism, Culture & Heritage (2004) *Bluenose*: *A Canadian Icon* [online], available: http://www.gov.ns.ca/nsarm/virtual/bluenose/ [accessed 13 April 2006].

Séguin, R. (2001) 'Quebec's Bernard Landry on the Maple Leaf: "A piece of red rag"', *The Globe and Mail*, 24 Jan, A1.

Singh, N. and Pereira, A. (2005) *The Culturally Customized Web Site: Customizing Web Sites for the Global Marketplace*, Oxford: Elsevier.

Smith, A. D. (1991) *National Identity*, Reno, Nevada: University of Nevada Press.

Yunker, J. (2003) *Beyond Borders: Web Globalization Strategies*, Indianapolis: New Riders.

# Localisation in The Netherlands: training and career opportunities

**Marcel Thelen[1], Han van de Staaij[1], Anne Klarenbeek[2]**
**[1]Department of Translation and Interpreting,**
**Maastricht School of International Communication,**
**Zuyd University Maastricht**
**[2]SDL Hengelo,**
**The Netherlands**
**m.m.g.j.thelen@hszuyd.nl, j.m.vandestaay@hszuyd.nl,aklarenbeek@sdl.com**

This article gives an overview of localisation in The Netherlands, both in education and industry. The discussion on education is further narrowed down to the area of training institutes that offer courses on the translation aspects of software localisation; the discussion on industry comprises the whole spectrum. On the education side, the article gives an overview of localisation courses offered in The Netherlands and the tools used in such courses. On the industry side the article gives an overview of the localisation market in The Netherlands, i.e. its players, the systems used and produced, etc. The discussion also focuses on the participation of industry in training. Finally, one of the authors, who is a graduate himself, briefly discusses his expectations and experiences.

**Keywords:** *Localisation, translation, The Netherlands, education, training, industry*

**Activities involved in localisation**

According to Esselink (2000:3) the following activities are involved in a localisation project: "(1) project management, (2) translation and engineering of software, (3) translation, engineering, and testing of online help or web content, (4) translation and desktop publishing (DTP) of documentation, (5) translation and assembling of multimedia or computer-based training components, and (6) functionality testing of localised software or web applications." These activities show that there are two distinct major roles to be played: engineering and translating. Each of these roles has a number of sub-roles. What these are can be derived from the more useful survey of the various aspects of localisation in Esselink (1998:6) where the people involved in a typical localisation project are given: "(1) Project Manager, (2) Translator, (3) Localisation Specialist/Senior Translator, (4), Proofreader/QA Specialist, (5) Localization Engineer, (6) Testing Engineer, and (7) Desktop Publisher."

For the education side of this article, the role of translating is understood to include — next to translating proper — project management and proofreading, and desktop publishing to a marginal extent only.

## 2. Education

### 2.1 Translation training institutes offering localisation courses

The Netherlands has six major translation training institutes, one of which is the Department of Translation and Interpreting, Maastricht School of International Communication, Zuyd University in Maastricht. Of these six institutes, the Department of Translation and Interpreting in Zuyd University is the oldest (founded in 1981 by HM Queen Beatrix) and is the only one offering courses on localisation (i.e. both translation with the help of CAT tools and localisation in the strict sense). The objective of these courses is not only to give students an introduction to the various aspects of localisation, but also to serve as the basis for actual work using a wide range of localisation and translation tools; in particular the courses strive to train the students to become skilled users of localisation tools and novice translation professionals (see [2.2] below).

### 2.2 Courses offered

The Department offers a four-year course in translation and interpreting at BA level. In *Year Two* of this course, there are four modules on translation that serve to introduce and instruct students in the use of CAT (Computer Assisted Translation) tools, namely Trados/SDLX. Almost all translation work to be done after the introductory modules has to be done with the help of these CAT tools. At the end of the second year, students work as junior translators and revisors (for three-and-a-half weeks full-time) for an in-house simulated translation bureau that is staffed and run by 4th year students under the supervision of a senior lecturer. During this period of working for the in-house translation bureau the students benefit greatly from using CAT tools (the use of CAT tools is made

compulsory so that the students' work in the simulated translation bureau mirrors as closely as possible that of a real-life translation bureau). As for the regular exercises and assignments throughout the second year, CAT tools are becoming household tools more and more. The same holds for the third and fourth years of the course.

In *Year Three*, there is an introductory module on localisation proper, i.e. an introduction to localisation tools and working with these tools. Although culture and institutions are already part of the regular language programmes, culture also plays an important part in this introductory module on localisation — focusing on the aspects of culture that are present in localisation and in particular the technical side of these aspects.

The bulk of the work on localisation takes place in *Year Four*, where there is a further specialisation in localisation, which takes the form of a project. This project can be practical (i.e. on the *actual* localisation of help files, software and documentation/manuals) or more theoretical (i.e. on the comparison of various localisation tools, the evaluation of a particular tool, etc.).

As can be seen, by the time of their graduation all students are skilled users of CAT tools and have enough knowledge to work with localisation tools, with a number of them even specialising further in localisation. What counts is that both groups are prepared and ready to work in translation bureaus (or start one themselves) and have the skills needed to work with CAT tools and localisation tools.

### 2.3 Tools used

During the second, third and fourth years of the course, a number of tools are taught and used. In *Year Two*, the CAT tools that are used are WordSmith, Trados MultiTerm, Trados Translator's Workbench (including the translation memory) and MultiTerm, TagEditor and WinAlign. There is also a course on HTML, albeit a basic course which reflects the ease with which students grasp this markup language. PASSOLO is covered in the introductory module on localisation in this year. The module contains a number of exercises on the use of this tool.

No further tools are introduced in Year 4, but this may well change in the near future.

### 2.4 Input from Industry

Input from industry takes a number of forms. Currently these are:

1 Guest lectures on CAT tools, localisation and the industry,
2 Third-year work placement — for 19 working weeks — in a country where the first foreign language of the student is the language of habitual use, and
3 Fourth-year work placement — for 10 working weeks — most often at a professional translation bureau in The Netherlands, although it is also possible to go abroad. Both the third-year and fourth-year work placements are compulsory.

This input from industry will be expanded greatly in the very near future, after which it will also include the following:

4 Collaboration in the development of teaching materials,
5 Participation in the more commonly termed skills laboratory (the in-house simulated translation bureau), participation in the more commonly termed 'learning company', and
6 Placements for lecturers.

Participation in the in-house simulated translation bureau entails sending translation and localisation jobs for further processing and giving feedback on the products delivered.

The 'learning company' is a new phenomenon where the Department of Translating and Interpreting actively searches the market for (innovating, if possible) real-life projects of varying durations for students to work on. On successful completion of such projects, students will earn credits. The idea behind this is knowledge circulation: industry gains from the work that is done by the students for the Department and the Department will be able to enhance its knowledge by closely cooperating with the industry experts.

### 2.5 Cooperation with Industry

The Department of Translation and Interpreting actively seeks to cooperate closely with industry. The type of training given at the Department is vocational by nature. Therefore, it is one of the main objectives of the Department to cooperate with industry in the areas of the curriculum and placements both for students and lecturers. As already described in Section 2.4, the Department is already rather successful in this respect (with plans in place to further

expand industry input in the near future). See also Section 2.6 for further discussion on this cooperation. The Department already liaises with both localisation producers and localisation translation companies in The Netherlands and abroad. One of the objectives for the future is to give the industry a greater role in the area of assessment.

### 2.6 Employability of graduates

The Department of Translation and Interpreting has a number of instruments to measure the employability of its graduates.

Firstly, the third-year and fourth-year placements are very important for factors in the employability of students. It happens very often that students doing their third-year placement are offered a job that will commence after their graduation, especially when doing their placement at a translation bureau. As for the fourth-year placements, it is a regular occurrence that placements lead directly to employment, with many placements being continued in the form of regular jobs. The Department works very closely with a number of renowned companies that offer such placements: SDL, Microsoft Ireland, Trados, various 'ordinary' translation bureaus in The Netherlands and abroad (mainly the UK), Lionbridge, Eclipse, RWS, and Philips Eindhoven. Also Medtronic (the world leader in medical technology) offers jobs to graduates at its translation and communication division in The Netherlands.

A second important instrument is the Department alumni scheme, central to which is an alumni website. More and more companies submit their vacancies for publication on this website and more and more graduates find jobs through this very same website.

*It is noteworthy that one of the former graduates from the Department of Translation and Interpreting, Maastricht School of International Communication, Zuyd University in Maastricht has now become one of the world's leading localisation authors, namely Bert Esselink.*

## 3. The industry

The localisation industry has been growing rapidly and continuously in The Netherlands since the 1980s when the world witnessed the first personal computer, for which various types of content needed translation. The localisation industry received a boost a few years later when the first translation memories appeared, making translation much cheaper, faster and more consistent. The third boost came from the emergence of the Internet. Suddenly, data was accessible anytime, to anyone, anywhere. This opened up the international market for literally everyone, creating a huge growth on the translation demand side. And the market is still growing. It is a market that is growing for every area of the industry; and one that is growing constantly for all areas (notwithstanding seasonal peaks, e.g. higher sale of electronic goods at Christmas).

Over the past two years, various factors have contributed further to this growth. On the IT and multimedia side, we have the upcoming Microsoft Office 2007 suite and Microsoft Windows Vista operating system, plus the rise in sales of home networking products, gaming products and domestic appliances. On the automotive/mechanical engineering side, new EU environmental directives have led to the development of new engines and vehicles, and more localisation work as a consequence. Another factor is that companies realise more and more that they will lose out on sales if they do not continue or start localising their products.

And let us not forget the joining of the most recent EU member states, which has led to an even greater demand for localisation, on top of the growing list of European directives which necessitate the localisation of all sorts of content. Lastly, within some agencies, the Dutch language has been added to the so-called FIGS list (French, Italian, German, Spanish), forming the tier 1 of languages for all localisation work that has priority for most clients of localisation companies. Officially, however, Dutch is still a B-language though it is coming closer to the FIGS list. This move augurs very positively for the localisation industry in The Netherlands as it indicates that the demand for localisation into Dutch is growing.

### 3.1 The market players

Since the acquisition of Trados by SDL and the acquisition of Bowne Global Solutions by Lionbridge (both in mid-2005), SDL and Lionbridge really are the two main localisation players in The Netherlands. These two market leaders make use of freelance translators and translation agencies of all sizes for their outsourcing needs. Many enterprises in various industries also run their own in-house translation departments, but regularly call on freelance translators and translation agencies when their internal resources are fully booked. There is a great shortage of translators in The Netherlands — in particular

in the localisation industry, thereby putting pressure on everyone at the supply end of the global information management chain. This shortage may be due to the growing demand for translation into Dutch (see the decision some agencies made to put Dutch on the FIGS list). The problem is that in a total population of 22 million Dutch-speaking citizens (Flemish included), there are not enough qualified translators. The shortage is also felt outside The Netherlands, e.g. at Microsoft in Dublin, Ireland where there is also a great need for native Dutch-speaking employees.

### 3.2 Expectations and experiences of a graduate

In this section, one of the authors, Anne Klarenbeek — who is a graduate himself — discusses briefly the expectations he had when he graduated and his experiences since then.

"Having graduated only four weeks earlier, I started working as an English-to-Dutch translator in August 2003. I quickly discovered that the pace was a lot higher than what we were used to at university. As I am working in a team that is specialised in the localisation of IT and multimedia content, I also noticed that, even though I had a greater than average knowledge of computer and networking hardware and software, I had a lot to catch up with. Personally, I found a great challenge (but also enjoyment and fulfilment) in jobs which require translators who are more skilled in 'transcreation' than translation — typically required for marketing pieces — and I noticed the same applies to newcomers who have joined the localisation industry over the past three years (albeit not everyone likes marketing pieces as much as others do). This area wasn't covered at university so I had to revert to my talents and the assistance of my co-workers.

My daily tasks also include file handling and resourcing. You could call it account management to some extent. My translation/review to account management ratio is around 70%–30%. This makes for a nicely varied pattern and a welcome change after a number of hours of concentrating on a piece of Help material or a user guide. The daily life of a localiser takes a lot of concentration and discipline and is often dynamic in the sense that one moment you are playing with words trying to sell a body groomer, and two hours later you are fixing the length of a handful of software strings, having just spent half an hour in-between outsourcing work, issuing purchase orders and answering translators' questions on the work they are helping you out with. There is never a dull moment if you like this kind of work".

## 4. The future

The prospects for localisation look promising in The Netherlands. Gradually, more training institutions are including localisation as a subject in their curricula, and in particular the Department of Translation and Interpreting of the Maastricht School of International Communication goes even further in that it is adjusting its curriculum to make it possible for the industry to actually take part in the training of prospective localisers (see Section 2.4). In addition to this, The Netherlands can boast to have the world leader in localisation, namely Lionbridge, and the world number two, SDL.

Lionbridge once started as a small Amsterdam-based localisation company named INK that gradually developed and expanded, changing its name once in a while until 1996 when the company became Lionbridge. Now the corporate headquarters are in Waltham, Massachusetts in the USA. The Amsterdam office is now a Lionbridge subsidiary.

SDL is originally a UK-based localisation company, with its headquarters in Maidenhead. Over the past years SDL expanded and took over other companies, among them Alpnet in 2001. Since then SDL has a subsidiary in The Netherlands (Hengelo). Both Lionbridge and SDL attract the world's greatest companies for localisation work: Lionbridge has the job of localising Microsoft's Vista and SDL has the job of localising Microsoft's Office 2007. Both companies are determined to strengthen their world position. All these elements can give localisation an even stronger position in The Netherlands.

**References**

Esselink, B. (1998) *A Practical guide to Software Localization,* John Benjamins Publishing Company, Amsterdam, The Netherlands/Philadelphia, USA.

Esselink, B. (2000) *A Practical guide to Software Localization,* John Benjamins Publishing Company, Amsterdam, The Netherlands/Philadelphia, USA

# GILC.org
# Software localisation by Translate.org.za

**Friedel Wolff**
**translate.org.za**
**South Africa**
**friedel@translate.org.za**

Prior to 2001, very little software had been localised for South Africa. The numerous languages in South Africa have been cited as problematic. Translate.org.za started to localise software for all eleven official South African languages and has been the catalyst and driving force in the South African localisation community. Initial work included a South African keyboard layout usable for all South African languages. Translations in all eleven official languages were recently completed for OpenOffice.org, Mozilla Firefox and Mozilla Thunderbird. These translation projects also resulted in the development of several localisation tools that simplify the localisation of these big projects and a web-based translation and translation management tool. An active volunteer community has already developed for one language and proprietary vendors have gradually been introducing localised products.

**Keywords:** *Localisation, South Africa, localisation tools, Afrikaans, minority languages*

South Africa has a long and complex history. It has seen some of the oldest hominids known to modern man, tribes migrating from the north, colonialism by the Netherlands and the United Kingdom, imported slaves, several wars, and the political turmoil of the 20th century.

Currently South Africa has eleven official languages. It is second only to India in terms of the number of official languages. However, the South African language diversity is quite small compared to other African countries. Linguistically, the ten non-English languages are categorised as follows: one West Germanic language (Afrikaans) and nine languages from the Bantu family. Of these nine, four belong to the Nguni language group, three to the Sotho language group, and two are separate languages (Tsonga and Venda). Several of these eleven languages are also spoken in neighbouring countries.

In some parts of the country English serves as a *lingua franca* — mainly in the cities and more so to the south-eastern part of the country. Afrikaans is more dominant in the western part of the country and in rural areas. The use of the other languages is, to varying extents, localised in certain areas.

Translate.org.za was started in 2001 by Dwayne Bailey to localise Free Software for South Africa. A few applications were available in Afrikaans but no software was available in the other local languages at that stage. Having eleven official languages were mostly of symbolic value when it came to computers and technology.

## 1. Input and display

There are relatively few technical problems with input and display for the eleven official languages. All use the Latin character set, with four languages using diacritics. The diacritics for three of these exist in some European languages. However, five characters exist that are unique to Venda. All of the extended characters had already been codified in Unicode.

No keyboard for South African languages has ever been developed. To this day, many resort to 'Alt-codes' in Windows to input the extended characters, or use application specific character insertion techniques, or simply do not use the correct characters any more. Neglecting the diacritics is truly problematic, as this greatly reduces the morphological wealth of the affected languages. In some cases people still manually insert diacritics after printing, thereby making it impossible to have perfectly correct electronic copies. The situation has probably been worsened by the fact that the non-English languages do not enjoy high status in business and the fact that the diacritics do not occur in all of the languages, and in some cases do not occur with great frequency either.

To rectify this, Translate.org.za developed a keyboard with which all languages of South Africa can be typed [1]. It could not be much different from the standard US layout, since that is what is ubiquitous in

South Africa, and many people need to use English regularly. It also had to be taken into account that

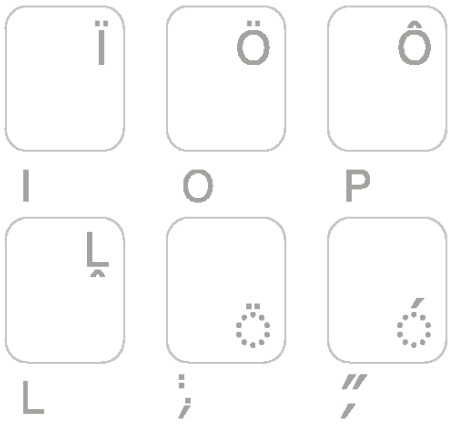



**Figure 1:** A sample of the keys for the right hand on the South African keyboard layouts. The dotted circles indicate dead keys for using the relevant diacritics.

many people would not use the extended characters often, or might not be adept typists.
Along with the development of the South African keyboard, the popular DejaVu fonts were extended with the Venda characters that were still missing.

## 2. Initial steps

As first projects, several smaller programmes were translated into the major languages of South Africa. Google South Africa was made available in four languages (unfortunately it is still not possible to limit searches to a specific South African language). The popular desktop environments, KDE and GNOME, were translated and valuable lessons were learnt: skilled translators were not always easy to come by and managing translation efforts for several languages proved to be a mammoth task. Many translators preferred to work with spreadsheets rather than with applications supporting the standard Portable Object (PO) files that are used to localise Free Software. Experience has shown that translators need more training than might be expected.

Work started on the development of some tools as part of the Translate Toolkit to do format conversion and to test the quality of translations automatically. Some of the automated checks included tests for consistent punctuation, spacing, variable use and XML tags. These tools proved to be invaluable for the review process, as many messages could be filtered out for review by a non-native speaker of the language.

## 3. Locales

Locales were added to the GNU C library and to OpenOffice.org to support all official languages. All locales are now available in the Common Locale Data Repository. Microsoft has supported Afrikaans and South African English for some time and they added three more South African locales in Windows XP SP2. Locales for most of the other languages have been assigned, but will only be part of future releases of Windows. Locales for two of the languages are still lacking [2] [3].

## 4. Big successes

In 2004 Translate.org.za released the first complete, localised Office productivity suite localised for South Africa by releasing OpenOffice.org 1.1 in four official languages [4]. As part of a large sponsorship Translate.org.za was able to extend the effort to update translations for OpenOffice.org 2.0 and include translations for all official languages [5]. The popular web browser and e-mail client from the Mozilla Foundation, Firefox and Thunderbird, were also translated into all official languages.

An important part of a fully localised office suite is a spell checker. Infrastructure was developed for the development of several spell checking systems and existing word lists were used to provide initial spell checkers. The complex morphology of the languages, especially for the languages in the Nguni group, offers severe challenges, especially for traditional UNIX spell checkers. The languages in the Nguni group are agglutinating languages, meaning that up to a whole sentence can be represented as a word. It is hoped that a future project would make it possible to extend Hunspell, the new checker used by OpenOffice.org, to provide support for the rich morphology of all official languages. Another difficulty with spell checkers is that they require good word lists. Often the best lists are to be obtained from lexicographers in the various languages. However, since the spell checkers are released as Free Software, it is very difficult to convince lexicography units — even though they are government supported — that this would not hamper their other commercial efforts in printed dictionaries.

## 5. Community building

Part of the effort of Translate.org.za was not only to provide localised software, but also to ignite the flame of community localisation projects and to build a culture of multilingualism and of using localised software. Because English is also an official lan-

guage, and English proficiency is relatively high amongst economically and technologically privileged, resistance to change has slowed the uptake of localised software. For some languages the community mailing lists are mostly dormant. The lack of translated teaching material has also been cited as an inhibitor for the adoption of localised software in training programmes.

To spark interest in software localisation, some *Translate@thons* (localisation sprints) were held with focus on specific languages. These events try to attract people to translate some software in a single day. The web based translation tool, Pootle, created by Translate.org.za, has proven invaluable for these events. Such events can attract a mixed crowd in terms of translation skill, technical skill, and true interest. While a small group usually creates quality localisations, large numbers make it very hard to achieve quality translations and probably serve better for creating awareness and interest. Most people still consider computers and electronics to be something inherently English.

However, interest in localised software is slowly on the increase. The major accomplishments are reported on in local internet news sites, and some interviews were held on national and community radio stations [6] [7] [8].

## 6. The Afrikaans localisation community

Afrikaans arguably sports the most successful localisation community among the endemic African languages. It has active mailing lists, coordinated terminology efforts and many localisation projects undertaken by community volunteers.

It was possible to compile a reasonably good spell checker from previously compiled word lists and this was improved by community members. More recently this work was also extended with hyphenation rules and data for the AutoCorrect feature of OpenOffice.org.

The success of the Afrikaans community has indicated that effort is often required by an individual to take initiative, or to coordinate willing helpers. Without leadership many projects are unlikely to be completed or to achieve good quality. Without an existing volunteer community, newcomers find it hard to become involved. An existing community provides means for newcomers to join, without them needing to provide the initial leadership.

## 7. The future

Although significant milestones have been reached, much remains to be done. The work on good spell checking for all languages was mentioned as an outstanding project; the complex morphology also affects development of an effective AutoCorrect functionality. Ideally grammar checking and thesauri should become feasible in the future. Collation specific for some of the languages should be considered, although the decision is a complex one, as powerful morphological analysis might be needed and could render such collating impractical.

Perhaps the most noticeable effect of the localisation work is to see how others have joined in. A few cellular phones are available in a few local languages, recently even with predictive text input for Afrikaans. A Zulu language interface pack for Windows XP was released in April 2006 and others would have followed soon thereafter [9]. Some of the Microsoft website is now also partially available in some South African languages.

Translate.org.za continues its work in the development of tools to simplify Open Source Software localisation; currently as part of the WordForge project. It is believed that this should afford even the smallest of marginalised languages a chance to efficiently manage their localisation projects at little or no monetary cost.

## 8. Conclusion

Despite common opinion to the contrary, we have proven that it is possible to localise software into all eleven official languages of South Africa. Only localising content for one of each of the language families is a common practice (a total of six languages), but localising for all eleven languages truly puts them on equal footing. We have also aimed to translate complete user interfaces, rather than only translating the most commonly used messages.

We have shown that Free and Open Source Software can act as a driver for localisation. Almost nobody had anything on the cards a few years ago, but since Translate.org.za has delivered, proprietary vendors have at least started doing lip service and stopped denying that there is a demand for localised products.

On the downside, we have found that dominance of English in the economic sphere makes it hard for localised software to be adopted, as those who would use it, do not want it to hamper their career prospects. Furthermore, the dominance of proprietary software suppresses localisation. The use of proprietary soft-

ware in government, education, the work place, etc. means that people don't necessarily even have the choice to use software in their mother tongue, even now that it exists.

We expect Free and Open Source Software to continue to dominate in localisation. As proprietary vendors follow this lead and create even more awareness, we expect more people to be drawn to community localisation efforts, where much greater depth and breadth is possible. Harnessing the power of communities empowers both the languages and the communities, and builds a culture of multilingualism.

## References

[1] Translate.org.za (2005) *South African keyboard* [online], available: http://translate.org.za/content/view/1526/51/ [accessed 11 July 2006]

[2] Microsoft Corporation (2006) *List of Locale ID (LCID) Values as Assigned by Microsoft* [online], available: http://www.microsoft.com/globaldev/reference/lcid-all.mspx [accessed 11 July 2006]

[3] Microsoft Corporation (2006) *Windows XP/Server 2003 - List of Locale IDs, Input Locale, and Language Collection* [online], available: http://www.microsoft.com/globaldev/reference/winxp/xp-lcid.mspx [accessed 11 July 2006]

[4] Translate.org.za (2004) *OpenOffice 1.1.2 released in Afrikaans, Zulu and Northern Sotho* [online], available: http://translate.org.za/content/view/1545/51/ [accessed 11 July 2006].

[5] Translate.org.za (2005) *OpenOffice.org Launch Celebration* [online], available: http://translate.org.za/content/view/1499/51/ [accessed 11 July 2006]

[6] Translate.org.za (2004) *Translate on the airwaves* [online], available: http://translate.org.za/content/view/1506/51/ [accessed 11 July 2006]

[7] Translate.org.za (2005) *SAfm with John Perlman* [online], available: http://translate.org.za/content/view/1496/51/ [accessed 11 July 2006]

[8] Translate.org.za (2004) *OpenOffice.org on Radio Pretoria* [online], available: http://translate.org.za/content/view/1563/51/ [accessed 11 July 2006]

[9] International Marketing Council of South Africa (2006) *Microsoft Windows in Zulu* [online], available: http://www.southafrica.info/public_services/citizens/your_rights/microsoft200406.htm [accessed 11 July 2006]

**Localisation Focus**
**The International Journal of Localisation**
**VOL. 5 (2006)**

## CONTENTS