Localisation Focus

THE INTERNATIONAL JOURNAL OF LOCALISATION

ISSN 1649-2358

SPECIAL CNGL EDITION

CNGL
Centre for Next Generation Localisation

sfi
science foundation ireland
fondúireacht eolaíochta éireann

The peer-reviewed and indexed localisation journal

VOL. 8 Issue 1

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: V811209

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

Opinions expressed by individual authors do not necessarily reflect those of the LRC or the editor.

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

It with great pleasure that I welcome you to this issue of Localisation Focus - The International Journal of Localisation, marking a milestone in the 15 year history of this publication. We are presenting a special issue focusing exclusively on the work of the largest-ever research project undertaken in localisation, that of the Centre for Next Generation Localisation (CNGL). The CNGL is a five-year collaborative project between four Irish universities and nine international industrial partners, co-financed by the Irish government's Science Foundation Ireland (SFI) and industry. Its value, in terms of both financial and intellectual investment, by far exceeds anything previously undertaken in localisation and affirms Ireland's world leading role in localisation. The Localisation Research Centre (LRC) at the University of Limerick (UL) is proud to lead the Next Generation Localisation research strand within the CNGL, complementing strands lead by researchers in Dublin City University (DCU), Trinity College Dublin (TCD) and University College Dublin (UCD).

In this volume, you will find contributions by Josef van Genabith, Director of the CNGL, on Next Generation Localisation, charting the map of the CNGL project; Lorcan Ryan, Dimitra Anastasiou and Yvonne Cleary on Using Content Development Guidelines to Reduce the Cost of Localising Digital Content, presenting a case for upstreaming localisation activities; David Lewis, Stephen Curran, Gavin Doherty, Kevin Feeney, Nikiforos Karamanis, Saturnino Luz, and John McAuley on Supporting Flexibility and Awareness in Localisation Workflows, describing approaches to smart workflow management and execution in localisation; Alexander O'Connor, Séamus Lawless, Dong Zhou, Gareth J. F. Jones and Vincent Wade on Applying Digital Content Management to Support Localisation, applying known digital content management approaches in a highly innovative way to localisation; Julie Carson-Berndsen, Harold Somers, Carl Vogel and Andy Way on Integrated Language Technology as part of Next Generation Localisation, describing the central contribution of core language technologies to increase efficiencies in localisation; and Ian R. O'Keeffe on Music Localisation: Active Music Content for Web Pages, reporting on his research into how music can be adapted to capture different moods and sentiments in different cultures. Last, but not least Chris Exton, Asanka Wasala, Jim Buckley and Reinhard Schäler present, in a special bonus article, Micro Crowdsourcing: A new Model for Software Localisation, reporting on the successful implementation of a demonstrator showcasing a ground-breaking, new model for software localisation.

Since its foundation in 1995, it has been one of our core objectives at the LRC to strongly promote, firmly anchor and decisively develop localisation and localisation-related issues in academic research by establishing the world's first dedicated academic research centre and teaching programme, by offering industry-sponsored academic research awards, by organising regular events such as the LRC Summer School and Annual Conference and by publishing the first peer-reviewed and indexed international journal for localisation. I believe that this volume demonstrates beyond any doubt that today we have achieved this objective - through the strong support of Science Foundation Ireland, the invaluable contributions and advice or our industrial partners and the excellent work of the more than 100 researchers collaborating in the CNGL.

Reinhard Schäler

Next Generation Localisation

Josef van Genabith
Centre for Next Generation Localisation
School of Computing
Dublin City University, Ireland
www.cngl.ie
josef@computing.dcu.ie

Abstract

Localisation is the process of adapting digital content to culture, locale and linguistic environment. Currently localisation is facing three massive challenges: volume, access and personalisation. In this paper we describe these challenges, outline core technologies, workflow and systems research tackling these challenges and introduce the Centre for Next Generation Localisation, an Industry-Academia partnership which carries out research to develop technologies and workflows for Next Generation Localisation.

Keywords: *Localisation, Machine Translation, Speech Technology, Adaptive Hypermedia, Information Retrieval, Crowd-Sourcing, Workflows, Software Engineering*

1. Introduction

In this paper we are beginning to chart the map of Next Generation Localisation, starting with the identification of three challenges to current state-of-the-art localisation operations and workflows. The challenges are volume, access and personalisation. We represent these challenges in terms of the localisation cube. We outline core technologies, including Machine Translation, Speech Technologies, Adaptive Hypermedia and Information Retrieval, to tackle the challenges in close combination with Crowd-Sourcing based approaches (where appropriate). Technologies (and Crowd-Sourcing) need to be integrated into flexible and adaptive workflows underpinned by standards and addressing the whole life-cycle of content development including design for internationalisation, localisation, distribution and reuse. Complex software systems instantiating adaptive workflows pose considerable software engineering challenges, many of which are currently underexplored. We provide an overview of the Centre for Next Generation Localisation, a large Industry-Academia partnership, which develops technologies and workflows for Next Generation Localisation.

2. Three Global Challenges to Localisation

Localisation is the process of adapting digital content to culture, locale and linguistic environment. Localisation brings products and services to markets that are otherwise inaccessible. Because of this, localisation is a core multiplier and value-adding component of the global software, services, manufacturing and content distribution industry. Currently, there are three massive challenges facing localisation: volume, access and personalisation.

- **Volume:** the amount of content that needs to be localised into ever more languages is growing steadily and massively outstrips current translation and localisation capacities. As a consequence, only a fraction of the content that needs to be localised is localised and usually only into a limited set of languages. Many business opportunities are missed and, what is more, lack of localisation contributes to the digital divide, with essential (e.g. health and hygiene) information, products and services unavailable in languages which currently do not promise return on investment (ROI) on localisation costs.

- **Access:** traditionally, localisation assumes print or full screen- and keyboard-based access to content. More recently, however, new and evolving generations of small devices (such as smart phones and PDAs) support on the move and instant access to digital content. Novel interaction modalities including speech-enabled access are currently not supported by localisation technologies. Traditional localisation workflows assume predictable, stable, corporate content and localisation is viewed as a well-managed and large-scale off-line process. Today, however, much digital content is perishable with frequent updates, including more and more user-generated content (user forums, blogs etc.). Instant access to such content requires a new breed of on-line localisation technologies.

- **Personalisation:** traditionally, localisation is coarse-grained according to generic notions of locales and linguistic environments (e.g. we localise for the Middle East). What is localised is information. Information is most valuable if adapted to personal and information requirements including task at hand, level of expertise, age-group and personal preferences and expectations. Traditional localisation needs to be overlaid and integrated with alternative and fine-grained personal information cutting across traditional notions of locale and linguistic environment. In terms of a slogan: the person is the ultimate locale.

Conceptually, the three challenges of volume, access and personalisation can be represented in terms of a localisation cube (Figure 1):

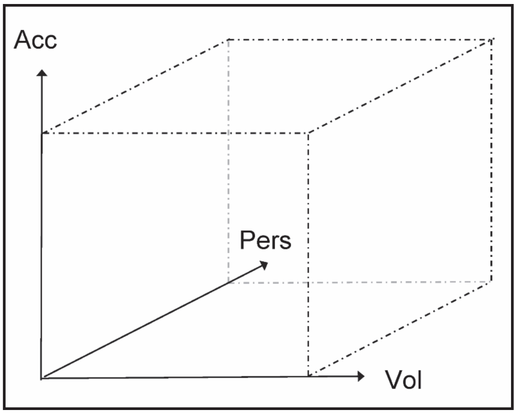


Figure 1. The Localisation Cube

The Localisation Cube captures a three-dimensional space defined by the coordinates of volume, access and personalisation. At the origin of the coordinates we have low volume, traditional print- and full screen- and keyboard-based access as well as low levels of personalisation, i.e. coarse-grained localisation according to generic notions of locale and linguistic environment. Moving from left to right along the volume-axis we approach today's large-volume localisation scenarios (as in large software localisation projects). Moving up the access-axis we go from off-line localisation processes with print- and full desktop- or laptop-based access to stable and predictable digital content into alternative access modalities (including speech-based interaction using smart devices) and instant on-line interaction with perishable and frequently updated digital content. Moving along the personalisation axis we move from coarse-grained localisation towards fully personalised access to and delivery of localised information. Current state-of-the-art localisation technologies by and large instantiate large and well-managed localisation workflows, targeting the lower, front-right part of the localisation cube (Figure 2), with large parts of the localisation cube remaining unaddressed.

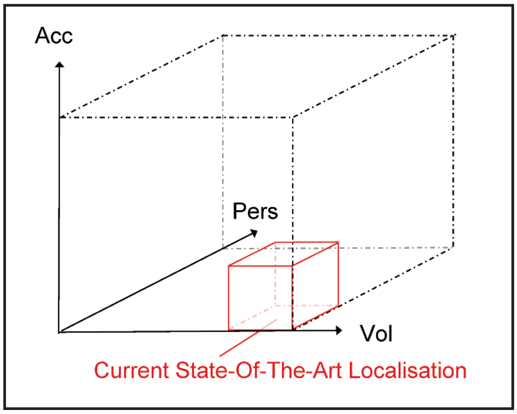


Figure 2. Current Localisation Technologies

The challenge is to develop next-generation localisation technologies and processes that allow us to address any point in the space defined by the localisation cube (Figure 1), at configurable speed and quality.

3. Addressing the Challenges I: Core Information Processing Technologies and Crowd-Sourcing

Information processing technologies are at the heart of Next Generation Localisation: Language Technologies (MT and Speech), automation and Crowd-Sourcing (and their combination) are the core approaches to address the challenges of volume and access; Adaptive Hypermedia and Information Retrieval (in a multilingual setting) jointly address the challenge presented by personalisation.

3.1 Machine Translation: Over the last ten years, machine translation technology (MT) has made substantial advances in quality and coverage with decreasing development costs: statistical MT (SMT) technologies (Koehn, 2009) are machine-learning based and trained on parallel text (human translations). As a consequence, MT systems can now be developed

- at a fraction of the time and cost of a traditional rule-based MT (RBMT) system (where the rules and lexical resources have to be hand-crafted by experts, a process that is labour intensive and expensive to scale to unrestricted text);

- for many more language pairs than was previously possible.

Furthermore, the training process involved in creating an SMT system can use data available from Translation Memory (TM) resources, a core technology used in current state-of-the-art localisation workflows. During operation, the resulting MT systems can be interfaced with TMs, fully leveraging previous translations, and terminology management resources (as available). Traditional RBMT systems can be composed with SMT systems in so-called automatic (or statistical) post-editing scenarios where an SMT system is trained on the output of an RBMT system, and then improves on the output of the RBMT system in a two-stage processing scenario (Tatsumi and Sun, 2009). This means that core resources and technologies from current state-of-the-art localisation workflows will not become obsolete but in fact provide essential components for next-generation workflows.

For high-quality translations involving human pre- and post-processing, deployment of state-of-the-art MT systems should result in productivity gains (addressing the volume challenge), changing traditional localisation workflows to include more substantial post-editing components. Furthermore, improved MT quality and availability for increased numbers of languages is crucial in enabling instant on-line access to digital content (without human post-processing) at configurable quality and speed and opens up the possibility of supporting languages and content beyond current ROI restrictions.

3.2 Crowd-Sourcing: crowd-sourcing involves volunteer groups in the translation (and hence localisation) process. This has been very successful in scenarios where volunteers are highly motivated. In the area of localisation, volunteers come in two types:

- user groups (usually not professional translators) effectively forming a fan-base of a particular product or service (e.g. social networking site users or cinema fans involved in the translation of movie scripts);

- professional translators (e.g. Translators without Borders) who work for free for a good cause for a Non-Governmental Organisation (NGO), which could not otherwise afford translation and localisation services.

Crowd-sourcing can

- drive down localisation costs;
- shorten turn-around times;
- produce culturally adept translation close to the pulse of the user group (this is a strong asset in the social networking type scenarios);
- produce training resources (parallel text) for SMT and TM applications.

From a corporate or NGO point of view, Crowd-Sourcing is attractive as the crowd works for free! However, and importantly, this does not mean that Crowd-Sourcing based localisation is for free: costs have to be factored in for crowd- and motivation-management and quality control. What is more, fan-based Crowd-Sourcing will not (in general) cover all text types: fans are highly motivated to translate user-facing documentation, text and web-pages etc. while they are not interested in translating (or indeed qualified for translating) technical or legal documentation (which often forms a large and important category in corporate localisation activities) and costs have to be factored in to cover these. With these provisos, Crowd-Sourcing will form an important factor in Next-Generation Localisation workflows. In particular, and in order to optimally support Crowd-Sourcing based efforts, it will be crucial to interface Crowd-Sourcing activities with full access to language technology-based translation aids including MT, TM and Terminology Management Support systems. Language technology, in turn, can benefit strongly from the parallel data (text) generated by the crowd: in particular machine-learning based approaches such as SMT will profit from the additional training resources provided by the crowd, providing improved MT back to the crowd in a virtuous circle (as already practised by Asia-Online and Google).

3.3 Speech Technologies: Automatic Speech Recognition (ASR) and Text-to-Speech Synthesis

(TSS) are mature technologies deployed in many applications (Holmes and Holmes, 2001). Current research challenges include speaker and language independence as well as robustness with respect to variations in (noisy) environments. Speech technologies are key to opening up alternative modalities for interacting with digital content (in addition to text- and screen-based interaction), particularly so in hands-busy and eyes-busy scenarios or for small portable electronic devices, such as smart phones and PDAs enabling access to digital content on the move. In some emerging economies, including India, small hand-held electronic devices have far greater penetration than desktop or laptop computers. In a multi-lingual environment speech-based content access and delivery needs to be tightly coupled with MT. In fact, Speech Technologies and SMT share basic underlying technologies: both are based on (variants of) the noisy channel model, and, in theory, both can benefit from information contributed by the other. To give just two examples: Speech Synthesis may make use of higher level information provided by MT systems, such as phrase or clause boundaries, or information about the part-of-speech (context) of a particular word or sequence of words. Machine Translation, in turn, may benefit from information provided by a Speech Synthesis component to select MT output (from a set of alternatives) that can be pronounced best.

3.4 Adaptive Hypermedia and Information Retrieval: in very abstract terms, what is localised is information. Information is most useful if adapted to the personal requirements, backgrounds and goals of a user. Adaptive Hypermedia (AH) (Chen and Magoulas, 2005) adapts information to user context, tasks and profiles. Traditionally, AH involves carefully handcrafted meta-data annotations, ontologies and content slicing and composition rules to ensure optimal content provision tailored to particular use scenarios and information requests, often in an e-learning setting. Typically, AH has operated over limited domains and carefully selected content repositories. By contrast, Information Retrieval (IR) (Singhal, 2001) has operated over vast amounts of diverse content (the Web) and is mostly machine-learning based. Personalised delivery of content is a core challenge for Next Generation Localisation. In order to meet this challenge, the robustness and scope of IR technologies has to be combined with the fine-grained and sophisticated content personalisation technologies developed in AH, all in a multi-lingual setting.

4 Addressing the Challenges II: Workflows and Complex Software Systems

Technologies are central to Next Generation Localisation. Technologies, however important, are not enough to solve the challenges of volume, access and personalisation: technologies need to be integrated into workflows and complex information technology-based software systems (the Next Generation Localisation Factory) supporting and instantiating these workflows.

It has often been remarked that in order to be maximally effective, localisation should not be treated as an afterthought. Indeed, upstream processes such as content development for localisation/internationalisation are crucial to successful localisation operations. Localisation workflows should consider the full life-cycle of (digital) content starting with content development, localisation proper and, finally, delivery and reuse. Content development may include controlled language, terminology, other meta-data annotation (do-not-translate etc.) and early reuse of content in development. Closely related to this are standards and their technical counterpart: interoperability. Standards (such as XLIFF, TMX, TBX) are required to ensure proper flow of information and meta-data in (complex) localisation workflows integrating content management, production, testing and delivery support systems, team (content developers, translators, localisation and software engineers, customer support etc.) and project management (costing, invoicing), crowd-sourcing (management, motivation, quality control) as well as translation, terminology, pre- and post-editing support systems. Interoperability ensures that different components are able to communicate on the system level through text and meta-data encoding standards, I/O interfaces, APIs and the like. Novel technologies such as Speech and Personalisation technologies need to be integrated with existing technologies and resources such as MT, TMs and Terminology Management Systems, and, for high quality translation output, with human pre- and post-editing and translators. Software systems supporting localisation workflows are complex and integrate sophisticated information processing technologies. To date, many software engineering implications involved in building, running and maintaining such systems are largely unexplored and it is not always clear what interface design will ensure optimal human-computer interaction with such systems.

5. The Centre for Next Generation Localisation

5.1 Basic Facts: The Centre for Next Generation Localisation (CNGL) is an Industry-Academia partnership developing Next Generation Localisation technologies. CNGL is funded jointly by the Science Foundation Ireland (SFI) and the industry partners under the Centre for Science, Engineering and Technology programme. The Centre involves about 120 people (with 50 PhD and 25 postdoctoral researchers) and the initial funding is for 2007-2012. The Centre is led by Dublin City University (DCU) and involves four university and nine industry partners. The academic partners are Trinity College Dublin (TCD), University College Dublin (UCD), the University of Limerick (UL) and DCU. The industry partners include large multi-national corporations (located in Ireland): IBM, Microsoft, Symantec and SDL, a large Japanese partner Dai Nippon Printing (DNP) and a number of SMEs operating in the localisation, translation and speech technology area: VistaTEC, Alchemy, Tràslan and SpeechStorm. CNGL has a flexible partnership and IP model with full and associate members, including the provision of contract research for fully external parties.

5.2 Structure of the Research Programme: in order to tackle the combined challenges of volume, access and personalisation, the CNGL research programme is structured as follows (Figure 3):

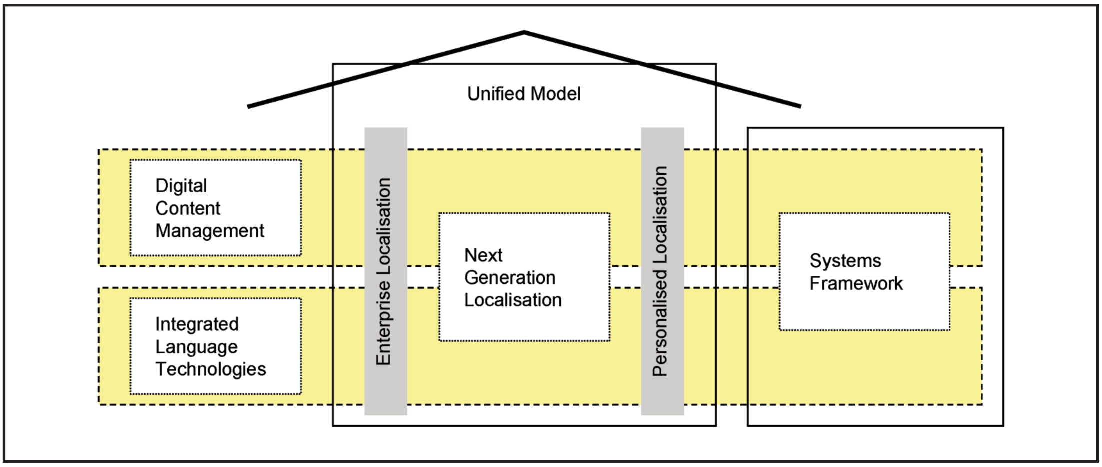

Unified Model
Digital Content Management
Integrated Language Technologies
Enterprise Localisation
Next Generation Localisation
Personalised Localisation
Systems Framework

Figure 3: Organisation of the CNGL Research Programme

The programme intertwines four research tracks: to a first approximation, two of them, Integrated Language Technologies (ILT) and Digital Content Management (DCM) are basic research tracks, and the remaining two, Next Generation Localisation (LOC) and Systems Framework (SF) are more applied, integrating research tracks.

5.2.1 ILT: ILT focuses on MT, with the aim of improving upon current MT technologies through integration of syntactic information in MT (in both SMT and example-based MT), the development of novel hybrid MT systems, automatic domain adaptation of MT, novel MT evaluation methods and investigating the impact of controlled language on MT. Furthermore, ILT has a Speech Technology component, closely intertwined with the MT research. Here the aim is to develop Speech Technologies that are less language dependent and can be adapted more easily to multilingual applications as well as the development of tightly coupled MT-Speech systems where the Speech system can profitably use information provided by the MT system and vice-versa. Finally, ILT has a Text Analytics component focusing on automatic annotation of localisation relevant meta-data, text classification (to e.g. support domain tuning of MT) and dependency annotation (to e.g. support syntax-enhanced MT).

5.2.2 DCM: DCM focuses on combining AH with IR technologies to support the CNGL personalisation agenda in a multilingual setting. In order to achieve its objectives, DCM focuses on the automatic acquisition of domain information and shallow subject ontologies from raw text, as manual construction of these is time consuming, prohibitively expensive and difficult to scale to large data sets and subject areas. Information queries are often the starting points of an interaction with digital content. DCM features a research component on query expansion and optimisation in multi-lingual contexts. Finally, content needs to be sliced and recomposed to deliver personalised information responses. DCM investigates novel methods based on insights from AH and IR for personalised multi-lingual information access and delivery.

5.2.3 LOC: the technological advances from ILT and DCM need to be integrated into the workflows of the Next Generation Localisation Factory. In order to achieve optimal integration, LOC researches the whole life-cycle of digital content, including content development and design for internationalisation. Standards are a crucial factor in achieving reusable and modular components in localisation workflows, and ensure that localisation-relevant information can be exploited optimally by those components. Sophisticated language and digital content management technologies need to be evaluated and integrated into workflows and combined with existing localisation technologies (such as TMs and Terminology Management Systems) and human pre- and post-processing. Finally, LOC will develop the blue-prints for the Next Generation Localisation Factory, which will be able to respond flexibly to localisation requirements addressing different points in the localisation cube (Figure 1) at configurable speed and quality.

5.2.4 SF: to date, software engineering aspects of complex language and digital content management technology based systems are underexplored. The Next Generation Localisation Factory will be highly modular and adaptive with easily, and on the fly, reconfigurable workflows. SF investigates rapid prototyping systems and designs supporting adaptive workflows, using web-based service architectures. User interfaces are a crucial component in such systems and novel interfaces need to be developed (to optimally support post-editing MT output). Finally, SF coordinates and implements the development of an evolution of CNGL demonstrator systems.

5.3 CNGL Demonstrator Systems: demonstrator systems are a core part of CNGL research activities. The demonstrators provide focal points for project cohesion and collaboration, combining technologies and teams from across CNGL. The demonstrators are essential for overall project evaluation and provide platforms for research and experimentation across the CNGL. They are also important showcases presenting CNGL technologies to the outside world. CNGL is developing an evolution of demonstrator systems instantiating important use scenarios in the space defined by the localisation cube (Figure 4):

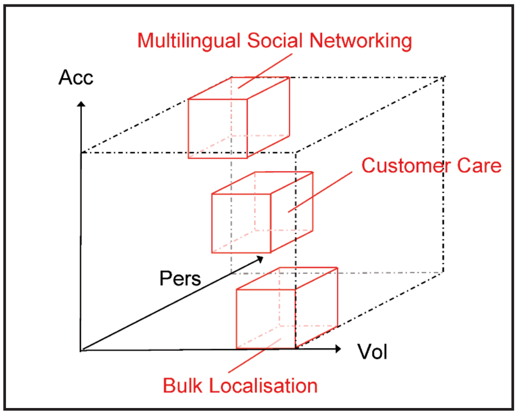


Figure 4: CNGL Demonstrator Systems Instantiating Use Scenarios in the Localisation Cube.

The Bulk Localisation scenario targets large volume localisation tasks, with and without human pre- and post-editing, familiar to current large localisation projects. The focus is on predictable corporate content, automation (MT) and the optimal integration of Crowd-Sourcing (where applicable) in an off-line localisation process with low levels of personalisation and standard print and full screen based access modalities. The Customer Care scenario focuses on users interacting with on-line and perishable digital corporate and user-based (product blogs) content, providing for frequent updates, speech-based access modalities (in addition to the more traditional modalities) and sophisticated levels of personalisation in real time interactions, without human pre- and post-processing interventions. Finally, the Social Networking Scenario focuses on user generated (in contrast to corporate) and highly perishable content prevalent on social networking sites, with high levels of personalisation and full use of all access modalities, using CNGL technologies to put networking sites in contact across linguistic barriers. CNGL demonstrator systems evolve according to a five year research programme (Figure 5):

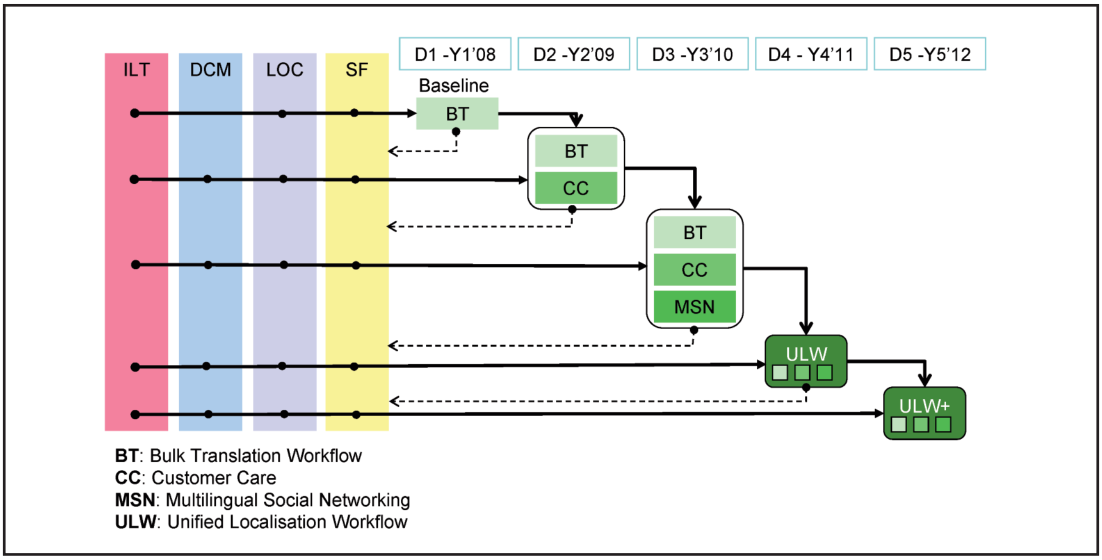

ILT
DCM
LOC
SF
D1 -Y1'08
D2 -Y2'09
D3 -Y3'10
D4 - Y4'11
D5 -Y5'12
Baseline
BT
BT
CC
BT
CC
MSN
ULW
ULW+
BT: Bulk Translation Workflow
CC: Customer Care
MSN: Multilingual Social Networking
ULW: Unified Localisation Workflow

Figure 5: Evolution of CNGL Demonstrator Systems

with a first base-line system instantiating a basic version of the Bulk scenario in year one, evolving into more sophisticated Bulk scenario systems in project years two and three. Project year two will see the first instantiations of Customer Care scenarios, evolving into more sophisticated systems in year three. Year three will see the first basic Social Networking applications. Years four and five will see the emergence of a Unified Workflow system, which can instantiate any point in the localisation cube (Figure 1) on demand with configurable quality and speed. In order to realise this trajectory, workflows need to be highly flexible, reconfigurable and adaptive: indeed, all demonstrator systems (starting with year one) are based on the same web services based architecture, sharing core components and development platforms.

6. Summary

This paper has charted Next Generation Localisation based on the identification of three challenges to localisation - volume, access and personalisation - and outlined technologies and workflow research tackling these challenges. The paper then presented the Centre for Next Generation Localisation (CNGL), an Industry-Academia partnership and its research programme designed to overcome the three challenges.

Acknowledgements

We gratefully acknowledge the support for CNGL from the Science Foundation Ireland (SFI), research grant 07/CE/I1142 under the Centre for Science, Engineering and Technology programme, and the industry partner contributions to CNGL from IBM, Microsoft, Symantec, Dai Nippon Printing, SDL, VistaTEC, Alchemy, Traslan and SpeechStorm. Any errors or inaccuracies in the presentation remain the responsibility of the author.

References

Chen, S., Magoulas, D. (2005) Adaptable and Adaptive Hypermedia, IRM Press

Holmes, J., Holmes, W. (2001) Speech Synthesis and Recognition, Taylor and Francis

Koehn, P. (2009) Statistical Machine Translation, Cambridge University Press (forthcoming)

Singhal, A. (2001), Modern Information Retrieval: A Brief Overview, in Bulletin of the IEEE Computer Society Technical Committee on Data Engineering 24 (4): 35-43.

Tatsumi, M. and Sun, Y. (2009) A Comparison of Statistical Post-Editing on Chinese and Japanese, in Localisation Focus, VOL.7 Issue 1, pp. 22-33

Using Content Development Guidelines to Reduce the Cost of Localising Digital Content

Lorcan Ryan[1], Dimitra Anastasiou[1], Yvonne Cleary[2]
[1]Centre for Next Generation Localisation
Localisation Research Centre,
[2] Department of Languages and Cultural Studies,
CSIS Department,
University of Limerick,
Limerick,
www.localisation.ie
lorcan.ryan@ul.ie; dimitra.anastasiou@ul.ie; yvonne.clery@ul.ie

Abstract

This paper examines how content development guidelines can reduce the cost of localising digital content. The growth of digital content is examined initially, and a basic taxonomy of enterprise and personal content is described. The demand for localised content is explained next, from international audiences, who demand content customised for their own particular locales. The costs, as well as cost-reducing strategies, involved in localisation process are also described. The cost-reducing strategy of internationalisation is focused on in particular, with the three core processes of authoring, enabling and testing explained in detail. The paper then describes how a Web 2.0 system called the Localisation Knowledge Repository (LKR) will be used to integrate content development guidelines into the internationalisation process. Finally, the benefits of the LKR are explained, including how the system facilitates the production of content for global audiences that is cheaper to translatable and requires less localisation testing.

Keywords: *Authoring, content development, digital content, internationalisation, localisation, pre-translation testing, technical communication, quality assurance, Web 2.0*

1. Physical and Digital Content

Content is a term that may be used to describe anything from a book, painting or video to an email, webpage or video game. Since the earliest cave paintings of the Stone Age, man has recorded ideas, information and opinions in a variety of physical written and illustrative media. Words, symbols and pictures have traditionally been captured on stone tablets, clay tokens, papyrus, vellum, canvas and paper over the years, and distributed as scrolls, magazines, books and paintings. Advances in technology, such as the invention of the earliest camera in the 1660s and first modern analog computer in 1930 (Encyclopaedia Britannica 2009), made it possible to store and communicate written and illustrative content in new formats such as photographic film, microfilm and magnetic tape. Physical content is the term we use to refer to words, graphics, music or video published in physical, non-digital formats such as canvas, paper or microfilm.

Media	Output Device	Format
Newspapers & Magazines	N/A	Paper, Microfilm
Books	N/A	Paper
Paper Documents	N/A	Paper
Maps	N/A	Paper
Video Reels	Video Players	Magnetic Tape
LPs & Music Cassettes	Turntables, Cassette Players	Vinyl, Magnetic Tape
Paintings	N/A	Canvas, Paper
Photographs	N/A	Microfilm, Photographic film

Table 1: Physical Content

Most content was published exclusively to these formats until the advent of the computer age, or information era, beginning in the 1980's.

1.1 What is Digital Content?

During the 20th century, the invention of the computer led to a new method of capturing data in computer files. We refer to information stored in this way as digital content. Digital content is content stored on hard drives or external storage media, published as a computer file or online, and accessed via a hardware device such as a computer, games console or mobile phone. It is viewed on display media such as computer monitors, televisions, mobile phone screens and eReaders, and shared via communication technologies such as the World Wide Web, email and Short Message Service (SMS).

A significant amount of content today is initially developed using software development, desktop publishing, help authoring, web design or graphic design software, and published as computer files. Any content published as a computer file is regarded as digital content, regardless of whether it is printed at a later stage or not. A written document such as a software user guide, for example, may be published initially as a .DOC, .PDF or .HTML file, and uploaded to a support website for customers to access. However, the same user guide may also be printed for packaging with the software product CD-ROM. Indeed, people not familiar with digital content may prefer to read certain materials in print format, such as books, photographs and timetables.

Media	Output Device	Example Formats
E-books	Computer, eReader	.LIT
Electronic Documents	Computer	.DOC, .PDF, .TXT, .HTML
Electronic Maps	Computer, GPS	.JPG, .PNG, .BMP, .GIF
Digital Video Files	Computer, DVD/Blu-ray Player, Mobile Phone, Games Console	.MPEG, .WMV, .AVI, .3GP
Streaming Videos	Computer, PDA, Mobile Phone, Games Console	.RM, .SWF
Digital Music Files	CD Player, Media Player, Computer, PDA, Mobile Phone	.MP3, .WMA, .WAV, .OGG
Streaming Music	Computer, PDA, Mobile Phone, Games Console	.IVR, .SMIL
CD/DVD/BD Software Applications	Computer	.EXE, .CLASS, .NET
Mobile Games	Mobile Phone	.JAR
Digital Photo Files	Digital Camera, Computer, PDA, Mobile Phone, Games Console	.JPG, .PNG, .BMP, .GIF
World Wide Web	Computer, PDA, Mobile Phone, Games Console	.HTML, .ASP, .PHP, .XML

Table 2: Examples of Digital Content

1.2 Advantages of Digital Content

Physical content is often converted to digital files with scanning equipment or conversion software; photographs, for example, may be scanned and saved as digital image files. Legacy content stored in physical media is generally digitised to avail of some of the following advantages associated with digital content (TidWiT 2009):

i. Storage: Digital content uses binary digits as the basic unit of information storage, rather than physical material such as paper or canvas. This allows digital content to be stored in massive quantities, with a palm-sized hard disk able to store tens of millions of pages of digital content. Digital content is also durable, and does not usually degrade as quickly as physical content (although some digital storage media may decompose over time).

ii. Classification: It can be easier to access and retrieve specific information in digital content than in its physical counterpart. Search facilities may be used to locate data in digital content, with hyperlinks also helping users to navigate to relevant chunks of information. Digital content is easy to restructure if necessary, and metadata may also be added to make it easier to find and retrieve information. Metadata also reduces the storage requirement for digital content, as the same file does not have to be stored in multiple locations just because it refers to several different topics.

iii. Accessibility: The connectivity afforded by the internet makes digital information much easier for publishers to distribute and for users to access. Content developers, for example, may now publish content online or email it directly to target audiences, rather than having to produce printed material. Digital content may also be made more accessible for disabled individuals by utilising assistive technologies such as screen reader software, Braille terminals, screen magnification software and speech processing applications.

iv. Publishing & Reproduction Costs: The cost of publishing and reproducing digital content is much lower for digital content than its physical content. Companies may create an electronic brochure and distribute it to thousands of customers simultaneously via email at relatively little time and cost. The same brochures would previously need to be printed, packed into envelopes and posted to each individual customer.

1.3 Factors Influencing the Growth of Digital Content

The volume of digital content produced today is significantly higher than ever before due to several factors, including the emergence of new electronic media, digitising of legacy content, user-generated content and corporate strategy. Each of these factors will be described in the following subsections.

1.3.1 Emergence of New Electronic Media

Since the initial rise in popularity of the personal computer in the early 1990s, electronic media have evolved into sophisticated handheld devices such as internet-enabled mobile phones, PDAs, eReaders and portable DVD players. A significant portion of content is now published in both its physical format and a multitude of additional digital formats for new electronic media. Books, for example, are generally published as both hardback and paperback printed copies; but may also be published as e-books, audio books on CD or MP3 and so on.

1.3.2 Digitising of Physical Content

A second factor for the huge increase in content production is the digitising of previously created content. Legacy content is usually converted to digital files with scanning equipment or conversion software; photographs, for example, may be scanned and saved as digital image files.

1.3.3 User-generated Content

The widespread adoption of the internet and the growth of Web 2.0 applications have created a new trend of users, rather than enterprises, publishing digital content. While digital content has traditionally been developed by professionals such as software developers and technical writers, millions of internet users are now creating and distributing digital information on a daily basis via instant messages, emails, forums and blogs. This combination of professional and social publishers means that the volume of digital content is continuing to increase at a considerable rate.

1.3.4 Corporate Strategy

Enterprises often prefer developing and distributing digital rather than physical content. Printing costs, for example, are reduced or eliminated by publishing technical support documentation online, rather than as paper-based product guides or user manuals. Distribution and packaging costs are also decreased by delivering software and other content to customers electronically, rather than sending them physical products. Digital content may also be distributed to users in a fraction of the time it takes to distribute traditional content.

1.4 The Growth Rate of Digital Content

If the world's rapidly expanding digital content was printed and bound into books it would form a stack that would stretch from Earth to Pluto 10 times (The Guardian 2009). As more people join the digital universe - through online access and internet-enabled mobile phones - the world's digital output is increasing at such a rate that those stacks of books are rising quicker than NASA's fastest space rocket (The Guardian 2009). The widespread adoption of home computers and usage of the World Wide Web in the 1990s accounted for the initial growth in production of digital content. This led to a boom in the computer hardware and software industry. According to the industry analyst INPUT, the total expenditure on software products in the US rose from $250 million in 1970, to $58 billion in 1995, to over $100 billion in 2000. After recovering from the "dotcom bust" of 2001, the domestic demand for packaged software in the US exceeded $126 billion in 2007 (International Trade Administration 2009).

Year	Total Expenditure on Software in the US
1970	$250 million
1995	$58 billion
2000	$100 billion
2007	$126 billion

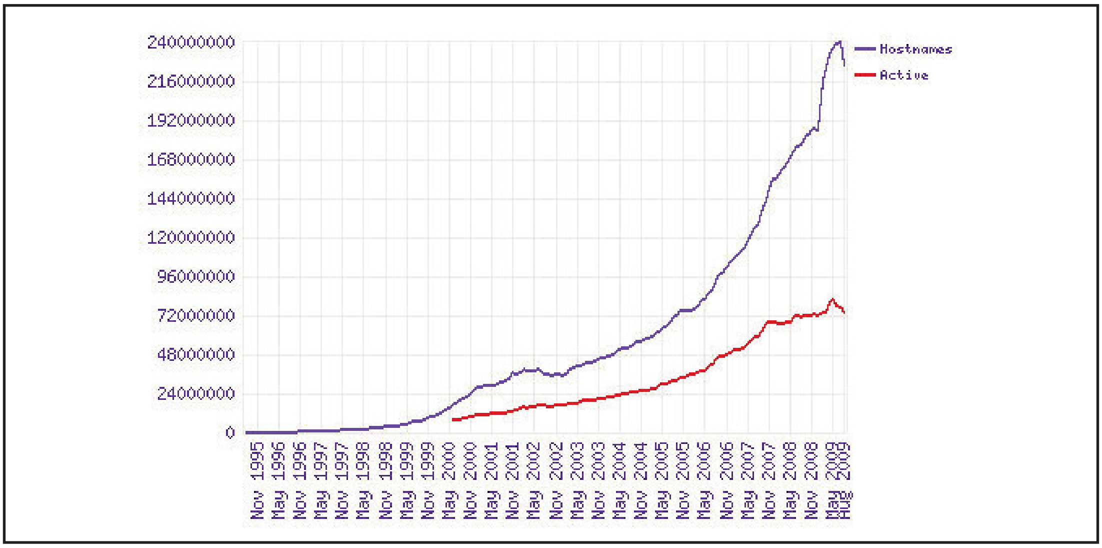
Table 3: Total Expenditure on Software in the US

Distribution of digital music also experienced rapid growth from the late 1990s on, with music downloads beginning to replace that of older media formats such as vinyl and cassette. The ratio of digital to analog sales in 2004 was roughly 1:99, but by 2007 it was roughly 1:9 (Recording Industry Association of America 2009). In 2008, physical album sales fell 20 percent to 362.6 million from 450.5 million units, while digital album sales rose 32 percent to a record 65.8 million units (Tahoe Daily Tribune 2009). Another indicator of the growth in digital music is the number of digital music file formats, which increased from less than 10 in 2003 to over 100 in 2007.

The amount of websites and the volume of online documentation has increased steadily in the last 15 years. There were just 18,000 websites when internet monitoring company Netcraft began keeping track in August of 1995. By June 2009 however, there were over 70 million active websites (Netcraft 2009) as shown in Figure 1.

As of May 2009, the world's digital content is estimated to be 487 billion gigabytes. The digital universe is expected to double in size over the next 18 months, according to the latest research from technology consultancy IDC and sponsored by IT firm EMC, fuelled by a rise in the number of mobile phones (The Guardian 2009).

Figure 1: Total Sites Across All Domains August 1995 - June 2009 (http://news.netcraft.com)



2. Enterprise & Personal Digital Content

Both digital and physical content may be sub-divided into further categories. The following section examines how digital content may be classified as either enterprise or personal according to who has published it.

2.1 Enterprise Content

Enterprise content is usually developed by professionals such as software developers, help authors, web designers, technical writers, graphic designers and technical engineers for commercial purposes. Software products, including desktop applications, firmware, video games, operating system software, business packages and software development kits are a significant component of enterprise content. The top five software vendors in 2008 were Microsoft, IBM, Oracle, HP, SAP and Symantec (Software Top 100 2009).

The quality of enterprise content is usually very high, as poorly written content could create negative perceptions in customers. Enterprise content, therefore, is usually:

- Published by organisations such as Sony Music Entertainment, EMI, Universal Music Group and Warner Music Group (digital music), Fox Entertainment, Paramount Motion Pictures Group, Sony Pictures Entertainment and Time Warner (digital video), Google, MSN and Reuters (online documentation)
- High volumes published by a relatively small number of content development professionals
- For commercial purposes
- Developed by professional authors
- Predictable content
- Static
- Published with professional tools such as software development kits, help authoring tools, web design packages and word processing solutions

2.2 Personal Content

Approximately 70% of the information in the digital universe is created by individuals rather than companies, and includes emails, photos, online banking transactions or postings on social networking sites (The Guardian 2009). This type of digital content, developed for social, non-commercial reasons, is known as personal content. Web 2.0 applications such as Google Wave, Blogger.com, YouTube, Facebook, World of Warcraft and Twitter (GoToWeb2.0 Web Applications Index 2009) enable users to instantly create, publish and share digital information via email, instant messaging, forums, blogs, social networking and massively multiplayer online games (MMOGs). Examples of users generating digital content include uploading videos to YouTube, publishing guitar tab on TabCrawler, sharing photographs on Picasa, sending emails via Gmail, posting blogs on Blogger.com, stating opinions on a personal website and creating "mod" games.

Personal content usually requires less quality control than enterprise content, as it is not published for commercial reasons and therefore doesn't normally undergo the rigorous testing and QA associated with enterprise content. Publishers of personal content may also use their own terminology and abbreviations. Online gamers, for example, use terms such as noob (a "newbie" or inexperienced gamer), leet (a sarcastic term referring to an "elite" player) and frag (to kill a computer game character), while SMS and instant messaging (IM) users regularly use abbreviations such as OMG (oh my god), GR8 (great) and LOL (laugh out loud).

Personal content, therefore, is usually:

- Developed and published by individuals
- Small in volume, published by a huge number of individual users
- For social or personal purposes
- Developed by social users
- Non-predictable, unique content
- Dynamic
- Usually published online

3. Localisation

Localisation is the process of adapting products, services and associated documentation so that they are understandable, acceptable and functional in target locales. Content is made understandable in locales by accurately translating it, acceptable by taking cultural differences into account, and functional by post-translation testing and editing. Inaccurate translations may cause confusion, lack of cultural sensitivity may cause offence, and content malfunctions may cause user frustration. Localisation should, therefore, encompass cultural awareness and technical expertise as well as the core activity of translation.

Enterprise Content (generated by organisations)	Personal Content (generated by individuals)
Software Applications	
Desktop applications	Homemade desktop applications
Video games	Homemade video games (WiiWare)
Help Systems	
Compiled help systems	User-driven forums
Electronic tutorials & wizards	
Web Content & Electronic Documentation	
Commercial websites	Homemade websites (Geocities)
Corporate emails & instant messaging	Personal emails & instant messaging
Online FAQs	Online customer support requests
Online help & live technical support (via IM)	User forums & threads (Boards.ie)
Corporate social networking (Linked In)	Personal social networking (Bebo, Facebook, Twitter, MySpace)
Commercial online documentation: - eBooks (Amazon.com) - eJournals (ACM Digital Library) - Online newspapers & eZines (IGN Magazine) - eMarketing (Google AdWords) - eLearning (e-learningforkids.org)	Personal online documentation: - Fan fiction ("fanfic") books - User blogs (Blogger.com) - Electronic fanzines - User reviews
Online games (World of Warcraft)	Teamspeak in online games (Half Life)
Commercial webinars (GoToWebinar)	Personal video communication (Skype)
RSS (Rich Site Summary) feeds	Wikis (www.wikipedia.org)
Copyrighted graphics (Fotosearch)	Personal photos (Flickr, Picasa)
Commercial videos (DVDs, WebTV etc.)	Homemade videos (Google Video)
Audio (CDs, iTunes, audio books, podcasts)	Homemade music (YouTube)
Commercial electronic documentation: - Maps (GoogleMaps) - User assistance (electronic manuals) - Marketing communications (electronic brochures) - Timetables, menus etc.	
Corporate text messaging (SMS)	Personal text messaging (SMS)

Table 4: Taxonomy of Digital Content

3.1 Why Localise?

Translation of physical content has been happening ever since the Hebrew Bible was translated into a Koine Greek version called the Septuagint between the 3rd and 1st centuries BC (Jobes and Silva 2001). Large scale translation of digital content, however, only began in the 1980s. Multinational corporations sought to increase international sales revenue by exporting translated software and documentation, but quickly realised that cultural and technical, as well as linguistic, issues required special attention. Therefore, to remain competitive in the global economy, organisations modified their exported products and services to give them the look and feel of locally-made products. This strategy is called localisation, and surveys show that nine out of ten businesses prefer to purchase products that have been adapted to their own language and market needs (Common Sense Advisory 2006).

Although digital content was initially translated by enterprises for business reasons, some organisations translate it for informative rather than commercial reasons. Parliaments, governments, councils and local authorities for example, translate digital documentation related to taxation and voting. The European Union (EU) produces legislation and documents of major public importance or interest in all 23 official EU languages (Europa Languages and Europe 2009). NGOs (not-for-profit organisations) including charities, foundations, social enterprises and humanitarian movements may also translate content to generate public awareness or collect donations for charitable causes. The World Health Organization (WHO) for example, publishes multilingual digital content with the aim of educating people about disease and promoting the general health of the world's population. Translators Without Borders also provide free translations to humanitarian organisations.

Internet users may also be motivated to volunteer translations for social reasons, such as prestige or the desire to share digital content with other users who speak their language. One Spanish user of Facebook for example, was responsible for translating almost 3 percent of the entire site for no other purpose than making it more accessible to other Spanish users (Facebook 2009). Facebook is also being translated into over 60 less widely spoken languages by its users, including Esperanto, Welsh and Afrikaans (Coyle 2009).

3.2 The Challenges of Localisation

There are several challenges involved in localising digital content; three of the most important can be classified under the headings of volume, cost, time and quality. The sheer volume of digital content, both enterprise and personal, coupled with the fact that 304 world languages have a million or more speakers (Ethnologue Languages of the World 2009), makes localising all of it a daunting prospect (Freij 2009). To furnish an example, the help system alone for Microsoft Office 2003 consists of over 700,000 words.

Commercial localisation is usually an expensive process involving investment in professionals and processes. Technologies such as machine translation systems aim to reduce the cost of localisation projects in the long run, but may require a significant initial investment to install and maintain. Localising personal content may be less expensive if crowdsourcing is used; i.e. the content is translated by users for little or no fee.

Multinational corporations often follow a strategy called "SimShip," or simultaneous launch and shipping of multilingual versions of their products in numerous locales. This puts enormous pressure on them to have localised versions of products ready in time for the source language market release. Some enterprise content may also require real-time translation such as live technical support or television subtitling. Users may request instant localisation of online communications with users in different locales such as emails, instant messages, forums, blogs and online gaming messages.

Enterprise content is usually developed for commercial reasons and either sold directly to customers (software products, digital music and so on) or produced to support the product offering (user assistance, marketing communications, legal documentation and so). High quality enterprise content, in both its original version and localised varieties, is essential therefore to maintain the image of the organisation (and product) with customers. Personal content is generally developed by users for social reasons, so the translation quality level is generally not as important.

3.3 Factors Affecting the Cost of Localisation

The cost of localising digital content is influenced by a number of factors, including the project scope, type of content and specific files involved. The scope of a localisation project is usually the most significant indicator of its cost. Scope can be measured in a number of ways including word counts, string counts, number of files and number of languages.

Enterprise content is typically subject to high quality translation, engineering and testing by localisation professionals. This type of project usually involves significant levels of investment in the appropriate processes, professionals and technologies. Translating a simple message on a social networking website on the other hand, where quality is not a central consideration, is generally a less expensive activity.

Some file types are more difficult to localise than others; flash files, for example, are time-consuming to prepare for translation. Image files with extensions such as JPEG and GIF may need to be sent to a graphic designer for localisation, as translators may not have the technical expertise to edit such files. These types of files add cost to localisation projects as they may be time-consuming or require additional specialised professionals to translate.

3.4 The Cost of Localisation

While the main expense involved in localising user-generated content is usually confined to the cost of translating it, large-scale enterprise content localisation projects may incur several different types of costs related to localisation processes, professionals and technologies. Planning, coordinating and implementing processes such as translation, quality assurance and project management may require significant investment of resources.

Localisation projects may require the skills of a diverse range of professionals from translators, technical writers, software developers and help authors to localisation engineers, proofreaders, testers and project managers. Hiring and coordinating these professionals represents another significant cost in commercial localisation projects.

Localisation technologies such as computer-aided translation (CAT) tools, software localisation suites and machine translation systems, often require enterprises to purchase licenses, install software and train users. These costs may be offset by the efficiency and productivity gains resulting from process automation however.

3.5 Reducing the Cost of Localisation

This paper focuses on reducing the costs associated with enterprise localisation projects. This type of localisation project usually involves digital content publishers sending source language content to a language service provider (or freelance translator) for translation. After proofreading and testing, the localised version of the digital content is published and distributed to international audiences. Companies attempt to reduce the costs associated with several aspects of enterprise content localisation projects, including the cost of localisation processes, professionals and technologies.

3.5.1 Localisation Processes

Organisations attempt to decrease the cost of the localisation projects by adopting cost-reducing strategies for component tasks such as translation, quality assurance and project management. Enterprises may attempt to reduce the cost of translation by using content development guidelines, best practices and standards to produce high quality digital content, which is easier to translate for both human translators and machine translation systems. Companies may also use technology to reduce the cost of translation. Machine translation systems for example, are used to automatically translate digital content, while translation memory tools reuse previously translated language strings into new projects.

Organisations may adopt a strategy called internationalisation to reduce the cost of localisation quality assurance (QA). Internationalisation is the process of generalising a product or document so that it can handle multiple languages and cultural conventions without the need for redesign. Internationalisation tactics such as writing for global audiences, enabling content for different locales and pre-translation testing ensure that digital content requires as little post-translation testing and localisation QA as possible. Enterprises may also use technologies such as localisation QA software, desktop publishing applications and software testing tools to automate QA procedures, and therefore reduce the costs associated with the process.

Companies may invest in project management software and workflow tools to reduce the cost of project planning, resource allocation and communication. Project managers may also experiment with new workflows to improve the efficiency and effectiveness of the overall localisation process.

3.5.2 Localisation Professionals

Companies may attempt to reduce the cost of hiring localisation professionals by using freelance translators or crowd-sourcing rather than procuring the more expensive services of language service providers (LSPs) or translation agencies. Localisation technologies such as machine translation systems may also be used to replace the human element and automate certain tasks.

3.5.3 Localisation Technologies

Enterprises attempt to reduce the cost of investing in localisation technologies in three important ways. Firstly, enterprises may choose to purchase inexpensive commercial technologies, even if the cheaper tools do not have as comprehensive a feature set as some of their more expensive alternatives. Localisation technology varies in price from CAT tools costing less than €1,000, to sophisticated content management systems which may cost several hundred thousands euros. A second method of reducing the cost of localisation technology is investing in open source rather than commercial tools. Open source software makes the source code available to the general public, and has relaxed or non-existent copyright restrictions. It is usually free to download and use, although costs may be incurred during installation, support and customisation. Finally, some companies simply develop their own proprietary localisation solutions, rather than use either commercial or open source tools. Although the research and development cost involved in this approach is initially quite steep, in the long run the enterprise does not have the expense of product upgrades, training or support contracts.

4. Internationalisation

This paper focuses on reducing the cost of localisation processes by implementing internationalisation guidelines using Web 2.0 technology. Internationalisation is the process of generalising a product or document so that it can handle multiple languages and cultural conventions without the need for redesign. The Localisation Industry Standard Association (LISA) (2009) suggests that it consists primarily of abstracting the functionality of a product away from any particular

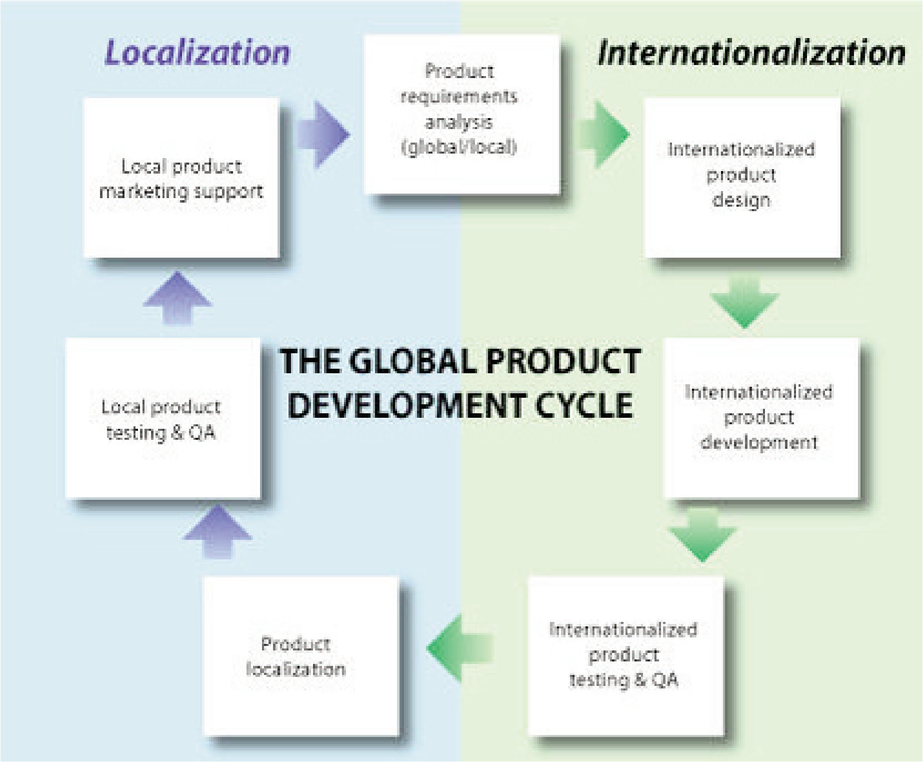

Localization
Internationalization
Product requirements analysis (global/local)
Internationalized product design
Internationalized product development
Internationalized product testing & QA
Product localization
Local product testing & QA
Local product marketing support
THE GLOBAL PRODUCT DEVELOPMENT CYCLE

Figure 2: The Global Product Development Cycle (Localization Industry Standards Association 2009 http://www.lisa.org/What-Is-Globalization.48.0.html)

culture, language or market so that support for specific markets and languages can be integrated easily. If a product has not been internationalised in advance, it can take twice as long and cost twice as much to localise, and such added expense may make it uneconomical to localise it at all (LISA 2009).

Internationalisation takes place at the level of program design and development, before the translation process (Figure 2). Enterprises should define which regions digital content will be distributed to before implementing the strategy, so that they are aware of the unique linguistic, cultural and technological nuances of each target locale.

Internationalisation and localisation are necessary due to variances in different locales, and to ensure that products, services and documentation are understandable and acceptable in different regions regardless of language or culture.

4.1 Diversity in Different Locales

Diversity in different locales can make content (which has not been prepared for translation) extremely costly and time-consuming to localise. We will describe three main types of diversity that distinguish locales, ie linguistic, cultural and technical diversity.

4.1.1 Linguistic

Linguistic diversity refers to variations in the language and writing conventions used in different locales. Sometimes even a single language may contain multiple regional varieties which should be considered in the content development process. Spanish, for example, with over 400 million speakers worldwide, has regional varieties such European Spanish, Castilian Spanish, Latin American Spanish, Standard Spanish, International Spanish and Neutral Spanish. Although all of these different forms are more or less understood in different locales, each variant has a unique vocabulary, pronoun usage and tense preference. "Costo," for example, is the Latin American Spanish term for cost, whereas in Spain, it refers to hashish in informal speech (Tek Translation 2009). Spelling variants may also exist for different countries where the same language is spoken; the word "localisation" in English (Ireland) for example, is spelled "localization" in English (U.S.). Another issue authors should consider is the meaning that a particular term in the source language has in other locales. General Motors discovered this to their detriment when promoting their Chevrolet Nova automobile in Spanish-speaking markets - "no va" means "doesn't go" in Spanish!

Writing conventions for currency, time, dates, weights and measurements may also differ depending on the locale. In France, for example, dates are written in the format day/month/year, but in Germany they are written day.month.year and in the Netherlands they are written day-month-year. In the United States however, dates are written month/day/year, and in Scandinavia (and some parts of Asia) are written year-month-day. The latter is defined by ISO 8601 as the international standard for writing dates (International Organization for Standardization 2009).

Considering issues such as weight measurement units during the content development process is also important. In 1983 an Air Canada Boeing 767 jet nicknamed the "Gimli Glider" completely ran out of fuel at 41,000 feet, halfway through its journey from Montreal to Edmonton. The pilot managed to crash-land the plane and nobody aboard was seriously injured, but an investigation was held immediately to deduce why this near-disaster occurred. The investigation found that the fuel requirement for the flight was set in metric units (20,000kg) while the local flight crew, who were used to calculating in imperial units, filled the plane with an incorrect fuel level (20,000lbs) which was insufficient for the flight (Aviation Safety Network 2009).

In a similar incident in 1998, the Mars Climate Orbiter was lost due to a navigation error when a subcontractor used imperial units (pound-seconds) instead of the metric units (newton-seconds) as specified by NASA (National Aeronautics and Space Administration 2009). Following this incident, NASA reverted back to using imperial units as their only system of measurement and continues to do so.

As well as writing conventions, other linguistic issues to consider are:

- Writing direction (Persian, Hebrew and Arabic are bi-directional languages running from right to left)
- Spacing rules (spaces are not used to separate words in Tibetan)
- Sorting order (if the content contains alphabetically-sorted lists, these will have to be reordered for each language)

4.2.2 Cultural

The culture of a particular locale consists of the shared attitudes, values, goals, and practices that characterise the area. No two locales share identical cultures, and it is important to capture this diversity during the content development process to avoid unintentionally irritating, offending or frustrating users in different locales with inappropriate content. Important cultural aspects that authors should consider are religious beliefs, political attitudes, colour associations, national holidays, sacred symbols, role of the family and so on.

The most famous example of publishing culturally inappropriate content in recent times occurred in 2005 when Danish newspaper Jyllands-Posten included 12 cartoons depicting the Islamic prophet Muhammed. This led to protests from many Muslims who felt the cartoons were Islamophobic, racist and blasphemous to people of the Muslim faith. Although the publishers of the cartoons maintained they were intended to be humorous and did not discriminate against Muslims, anger remained over the offensive images. Consumer boycotts of Danish products were organised, several Danish embassies were attacked and death threats were issued to the illustrators of the cartoons. Although the Danish Prime Minister apologised for any offense caused by the cartoons, tensions remained between Denmark and some sections of the Islamic world.

As well as having knowledge of political and religious beliefs, it is also useful for content authors to understand the subtle nuances of target locales such as colour associations. Purple, for example, is associated with royalty and wisdom in western cultures, but represents mourning in some Asian countries. By incorporating cultural awareness into the authoring process, the risk of publishing content that is offensive to users in different locales is minimised.

4.3.3 Technical

In addition to linguistic and cultural diversity, content authors should also be aware of technical variances in different locales. Character encoding is one of the most important technical issues to consider when developing digital content for global audiences. A character encoding system consists of a code that pairs a sequence of characters from a given character set with a sequence of natural numbers. Authors should ensure that the system being used supports all characters in the language in which the content is being developed.

Another important issue to consider is whether keyboard shortcuts and hotkey combinations will work correctly in different locales. One must also be aware of the technical infrastructure, mobile devices and so on used in different locales. Online content, for example, should function correctly in different locales regardless of the operating system or browser. Software applications may have additional technical concerns such as string concatenation or hard-coded strings.

LISA estimates that content developed without consideration for the linguistic, cultural and technical variances in different locales takes twice as long and costs twice as much to localise (LISA 2003). This paper proposes incorporating guidelines into internationalisation strategies to ensure the development of high quality digital content that takes into account the variances between different locales. Before examining how these may be incorporated to reduce the cost of localisation, it is necessary to examine each of the activities involved in the internationalisation process.

4.2 Author-Enable-Test Strategy

The internationalisation of digital content consists of several key tasks, the most significant of which are authoring, enabling and testing. The implementation of these tasks during the digital content development process is referred to as an AET Strategy in this paper (Table 5).

4.2.1 Authoring

Authoring is the process of writing or constructing a document or system. Authoring of source language content may also be considered to be the first step in the localisation process, with its quality level having a significant impact on the cost of translation and localisation QA. The main objective of the authoring process is to develop source language digital content that is usable and translatable. Usable content increases satisfaction among local users, while translatable content decreases the cost of localising the content for international users.

Content authoring research is based on the academic field of technical communication, with a focus on linguistics, content development guidelines, cultural research and technical writing standards. Specific professionals are dedicated to authoring different types of enterprise content; software applications, for example, are generally authored by software developers while digital documentation is usually generated by help authors, web designers and technical writers. These professionals typically use

	Author	Enable	Test
Objective	Develop clear, well-written content for international use	Prepare digital content for international use and localisation, so that it is easier to translate, engineer and test	Check digital content for internationalisation prior to translation
Research Area	Technical Communication	Internationalisation	Software Testing
Specific Considerations	Technical writing, linguistics, authoring guidelines, cultural research, technical writing standards	Software engineering, web design, help authoring, document design	Proof-reading, test compiling, desktop publishing, software testing, website validation
Tasks	Developing digital content using authoring tools, corporate glossaries, style sheets, controlled language techniques	Using Unicode character encoding, designing software user interfaces and document layouts to accommodate for expansion of translated text, setting maximum limits for string lengths to avoid layout problems, eliminate string concatenation, separate content and functionality.	Linguistic testing (proof-reading, consistency checking, checking language formatting, verifying sentence lengths), cosmetic testing (visual inspection, pseudo-translation, verifying character encoding) and functionality testing
Post-Translation Consequences of Non-Implementation	Inaccurate translations due to unclear sentences, poor grammar and punctuation, and inconsistent terminology in source language content	Poor layout (clipped text, overlapping controls) or characters not displaying correctly due to source language content not being properly enabled for localisation	Non-functioning (or malfunctioning) digital content with inaccurate translations or cosmetic issues due to source language content

Table 5: Components of an AET Strategy

authoring tools such as Microsoft Word, Adobe Framemaker, Adobe Dreamweaver and Madcap Flare to develop digital content. These tools allow authors to check the linguistic quality of the content they are creating, although issues specific to localisation (such as character encoding) may not always be included in these validation checks.

Content development guidelines published by researchers, organisations or professional authors assist the authoring process. These guidelines usually develop from academic research, industry best practices and proposed standards. Authors also use controlled natural languages (CNLs) to prepare content for localisation by deliberately restricting the grammar and vocabulary to reduce or eliminate ambiguity and complexity. CNLs, such as Simplified English and E-Prime, are adopted by some organisations to increase the readability and translatability of digital content. Simplified English offers a carefully limited and standardized subset of English designed to reduce ambiguity, make human translation easier and facilitate machine translation. E-Prime attempts to generate similar benefits by eliminating all forms of the verb to be: "be", "is", "am", "are", "was", "were", "been" and "being" (and their contractions). Authors may also use corporate dictionaries, glossaries and terminology databases to improve the consistency of content developed. Some organisations issue style sheets to authors to ensure the content they develop is as consistent as possible.

It is also essential for authors to consider cultural nuances when developing digital content for international audiences. These cultural aspects may be collectively examined by constructing a PESTEL analysis for each locale which includes:

- Political Considerations (government type, political history)
- Economic Considerations (purchasing power, standard of living)
- Socio-cultural Considerations (language, religion, attitudes, customs, colours, myths, symbols, fashion, education, role of the family)
- Technological Considerations (technical expertise, computer hardware)
- Environmental Considerations (natural resources, attitude to the environment)
- Legal Considerations (legal implications such as financial regulations & FDA requirements)

Poorly authored digital content, therefore, may contain incorrect grammar and punctuation, unclear language or inconsistent terminology. Time, dates, currency and measurement units may be used inconsistently throughout the content. References to religion, politics and symbols also have the potential to be offensive to users in different locales. Problems like these not only reduce the quality of the source language content for local users, but also make it more difficult to translate for international users.

4.2.2 Enabling

Enabling is the process of preparing digital content at a technical level so that it can handle multiple languages. Translating digital content that is not properly enabled for localisation may result in errors such as poor layout, clipped text and overlapping controls, characters not displaying correctly or international keyboards not working correctly with the content. Enabling aims to prepare content for localisation so that the process of locating and rectifying post-translation errors is less costly and time-consuming. Some of the main tasks involved in enabling digital content for localisation are:

- Using the Unicode character encoding standard to ensure that all international character sets are supported in the content, including bi-directional scripts such as Arabic and Hebrew
- Designing software user interfaces and document layouts to accommodate the expansion of translated text
- Setting maximum limits for string lengths to avoid layout problems

Software engineers, web developers or dedicated internationalisation professionals are usually responsible for enabling digital content for localisation at a technical level. These professionals use content development tools (software development kits, help authoring software, web design packages, desktop publishing applications) or dedicated internationalisation tools (such as Globalyzer Diagnostics) to assist them in completing enabling tasks.

4.2.3 Testing

Testing, as a part of an AET internationalisation strategy, refers to checking the linguistic, cosmetic and functional quality of digital content, prior to translating it into different languages. Linguistic quality refers to how readable and translatable the content is, as well as how culturally appropriate it is for different locales. Linguistic quality is usually checked by a technical writer or editor. The main tasks involved in checking the linguistic quality of the source language content are:

- Proof-reading to ensure clarity of expression, grammar and punctuation
- Consistency checking to ensure adherence to corporate glossaries or termbases
- Checking cultural conventions such as time, date, weight and measurement formatting
- Verifying word and sentence lengths are appropriate for human or machine translation
- Checking for: hardcoded strings; encoding; concatenation

Enterprises using Controlled Natural Languages (CNL), style sheets, corporate terminology databases or authoring tools usually spend far less time testing linguistic quality before translation than organisations that have no structured global content development process. Digital content that has not been tested for linguistic quality prior to translation, is more time-consuming for human translators to translate, and more difficult for machine translation systems to process.

Cosmetic quality refers to how visually consistent and aesthetically pleasing digital content is; this is usually checked by a technical writer, help author, software engineer or desktop publishing specialist. The main tasks involved in testing the cosmetic quality of digital content, before translation, are:

- Visual inspection of software user interfaces and document layouts to ensure there is room for text to expand after translation
- Pseudo-translation to preview the impact that translation is likely to have on the source language digital content
- Verifying that the character encoding is appropriate for display in different locales

Testing the international functionality of digital content is essential to ensure a high standard of quality, regardless of whether the content will be localised or not. Digital content published with functionality errors will not only have to have these rectified in the source language version, but also in each localised version. Test engineers are usually responsible for testing the functionality of source language digital content, completing tasks such as:

- Test compiling, virus scanning and checking functionality of software applications
- Checking the operability hyperlinks and search boxes of online documentation

5. The Localisation Knowledge Repository (LKR)

The Localisation Knowledge Repository (LKR) is proposed as a method of incorporating content development guidelines into the AET process, with the aim of making digital content more translatable and less expensive to localise. The LKR (currently in development) consists of three distinct sections, the Digital Library, Test Area and Virtual Community. Each area of the LKR is based on user requirements generated from active PhD research projects running in the Localisation Research Centre (LRC) located in the University of Limerick as part of the Centre for Next Generation Localisation (CNGL) project.

5.1 LKR Digital Library

The LKR digital library is an online repository of content development guidelines, cultural guidelines and relevant industry standards. The Digital Library is initially populated with content development data generated through primary and secondary research. The LKR also incorporates a feedback loop enabling users to upload content development and cultural guidelines. Any uploaded guidelines are reviewed by a moderator before being published to the Digital Library in order to maintain the quality of the LKR system. Another means of ensuring the usefulness and relevance of the guidelines is by providing users with a rating system where they may publish a comment about each guideline and rate it on a five-point scale.

5.1.1 Content Development Guidelines

Content development guidelines are instructions, principles and best practices for content developers writing digital content for international audiences. These guidelines are compiled from primary research, existing literature and industry best practices. Most of the guidelines are sourced from the academic field of technical writing, but other relevant areas include internationalisation, web design, help authoring, software development, document design, linguistics, controlled natural languages (CNLs) and terminology management. Additional secondary research is generated from company reports and case studies. Primary research is generated from interviews, focus groups, surveys, and usability testing with technical writers, help authors and document developers.

Content guidelines stored in the LKR Digital Library are classified into five categories for easy access:

- Content (language style, voice and tone, punctuation and grammar, sentence length, terminology, graphics)
- Presentation (font type and size, use of colour, blank space, page layout)
- Navigation (table of content, navigation maps, reading sequences, search boxes, indexes)
- Accessibility (access medium, features for disabled users)
- Other Issues (functionality on different hardware, operating systems and web browsers)

5.1.2 Cultural Guidelines

The second section of the Digital Library enables users to define a particular locale and access a selection of significant cultural considerations associated with it. Only cultural research significant to digital content development and localisation is stored in the Digital Library, as creating a repository detailing cultural aspects of every world locale would make the system unwieldy and unmanageable.

Relevant guidelines, therefore, are extracted from cultural research conducted on each locale, and published in the cultural guidelines section of the LKR Digital Library. The cultural guidelines are classified according to a PESTEL analysis (see 4.2.1).

5.1.3 Content Development Standards

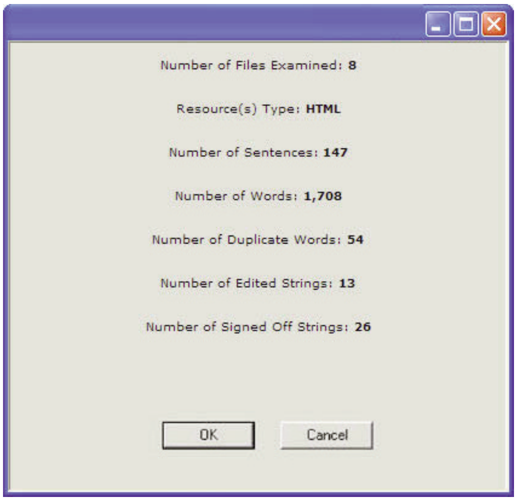
A standard is an established norm or requirement outlining necessary criteria, methods, processes or practices. Standards may be developed by corporations, trade unions, industry associations or dedicated organisations. Several organisations develop and publish standards relevant to content development and localisation including the LISA, Organization for the Advancement of Structured Information Standards (OASIS), World Wide Web Consortium (W3C), International Organization for Standardization (ISO) and the Unicode Consortium. Relevant content development and localisation standards are included in the LKR Digital Library including:

- Localisation standards (XML Localization Interchange File Format (XLIFF), Segmentation Rules eXchange (SRX) and Translation Memory eXchange (TMX))
- Time and date formatting standards (ISO 8601)
- Usability standards (ISO 9241)
- Character encoding standards (Unicode)
- Web standards (ISO 8879:1986 SGML)

These standards help content developers publish consistent, high quality digital content.

5.2 LKR Test Area

The Test Area enables content developers to use the LKR system as a test bed to check the quality of digital content. The user accesses the LKR Test Area, opens a new project and specifies which files to check. The LKR then parses these files, showing the user a list of all the language strings in the file. After all the relevant files have been "checked-in" (i.e. a copies of the source files are uploaded to the LKR website), users may select a View Project Statistics option to display attributes about the project such as number of files, number of sentences, number of words, number of duplicate words, number of unedited strings, number of edited strings and number of signed off strings. A Generate Report option enables users to create project reports containing vital project statistics.



Number of Files Examined: 8

Resource(s) Type: HTML

Number of Sentences: 147

Number of Words: 1,708

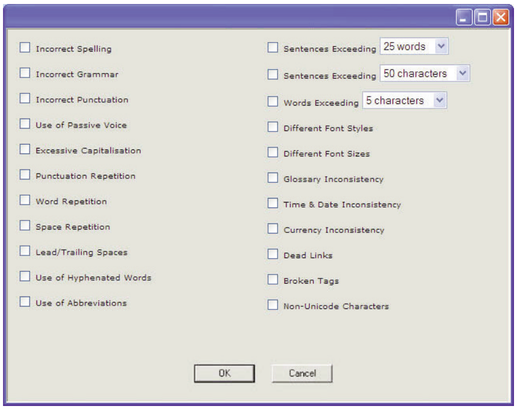
Number of Duplicate Words: 54

Number of Edited Strings: 13

Number of Signed Off Strings: 26

Figure 3: LKR View Project Statistics dialog box

Users may also select a Check Content option to check the project files for a predefined list of quality issues, from poor spelling and grammar to repetition, inconsistency and broken tags. The result of the search is displayed in a list for the user, who has the ability to rectify any issues in an editing window.



□ Incorrect Spelling
□ Incorrect Grammar
□ Incorrect Punctuation
□ Use of Passive Voice
□ Excessive Capitalisation
□ Punctuation Repetition
□ Word Repetition
□ Space Repetition
□ Lead/Trailing Spaces
□ Use of Hyphenated Words
□ Use of Abbreviations

□ Sentences Exceeding 25 words
□ Sentences Exceeding 50 characters
□ Words Exceeding 5 characters
□ Different Font Styles
□ Different Font Sizes
□ Glossary Inconsistency
□ Time & Date Inconsistency
□ Currency Inconsistency
□ Dead Links
□ Broken Tags
□ Non-Unicode Characters

Figure 4: LKR Check Content dialog box

Once the relevant changes have been made, users can save the project and select the Export File option to generate an edited version of the original source file (in the same file format). Users can also choose the Pseudo-Localise option to export a file with a predefined level of text expansion to simulate the impact that translation might have on the source files.

5.3 LKR Virtual Community

The final component of the LKR is the Virtual Community. It consists of three sections:

- Forums Area (where LKR users can share ideas and opinions)
- Resources Area (where LKR users may upload and download resources such as style sheets, glossaries, termbases, corporate dictionaries and relevant multimedia files)
- Connect Area (where LKR users may contact other users by email or instant messaging)

5.4 Web 2.0 Features of the LKR

The LKR is an online repository that uses a selection of Web 2.0 features (O'Reilly Media 2009):

- User-generated content (users upload their own guidelines and resources to share with others)
- Crowdsourcing (users maintain the quantity of LKR data by uploading guidelines and resources, and maintain its quality by rating the guidelines and commenting on them (Trieloff 2007))
- Social networking (users express ideas and opinions via the Forums Section in the Virtual Community, and connect to other content developers via email and instant messaging)
- Web applications (the LKR Test Area is a functional system operating as a web, rather than a desktop, application)
- Customisation (when users create a LKR log-in profile, they define the type of content they develop, locales they work with and so on, so that the data displayed to them by the LKR system is as useful and relevant as possible)

5.5 Using the LKR to Integrate Content Development Guidelines into the AET Process

Integrating content development guidelines into the AET process reduces the cost of localising digital content by making it more translatable. The LKR system enables content developers to integrate these guidelines into their workflows by providing a free, accessible database of the most up-to-date content development instructions, based on academic research and industry best practices. It also enables them to check-in the files they are working on, and conduct a customised quality check in the LKR Test Area. Users may customise the type of quality checks run in the LKR Test Area; a software developer, for example, may check for overlapping controls in a software user interface, while a technical writer may check for incorrect spelling or inconsistent terminology. The LKR, therefore, assists professionals in incorporating content development guidelines into the AET process in the following ways:

- **Authoring:** The LKR Digital Library gives content developers access to a vast repository of content development guidelines, cultural guidelines and relevant industry standards. They may also download glossaries, termbases and style sheets from the library, and contact other content developers for advice via the LKR Virtual Community.
- **Enabling:** The LKR Test Area supports the process of enabling digital content for localisation. LKR users, for example, may set maximum character limits on language strings to reduce the risk of their post-translation expansion corrupting the layout of the content. The LKR Digital Library also contains guidelines on internationalisation and enabling content for international audiences.
- **Testing:** The LKR Test Area enables users to perform automatic linguistic testing (consistency checking, checking language formatting, verifying sentence lengths), cosmetic testing (pseudo-translation) and functionality testing (exporting edited versions of project files to check that they function correctly).

The LKR may be useful in the following scenarios:

- A web developer creating or updating web content
- A technical writer developing an online manual
- A help author designing a web help system
- A content specialist writing an e-learning course
- A marketing executive announcing a new product in a HTML email flyer

5.6 Benefits of the LKR

The LKR, as proposed, has the potential to deliver a number of benefits to its users:

- Reduce the cost of developing and localising digital content for global audiences
- Increase the quality of source language enterprise content for local user
- Improve the translatability of enterprise content (both for human translators and machine translation technologies)
- Provide a free system that marries industrial practice with academic research, and is maintained by its user base
- Preserve minority languages by making it easier to translate from and into more widely-spoken

and commercially viable languages
- Bridge the digital divide by making it easier to develop digital content suitable for multiple locales (not just developed economies)
- Create an international network of content developers to share knowledge and opinions
- Inject cultural considerations into the localisation process (some view it as a purely technical activity)
- Provide an alternative to commercial desktop applications sold for profit (the LKR will be a free web-based application based on academic research, industry standards and enterprise best practice; incorporating Web 2.0 features such as customisation, social networking and crowdsourcing).

These potential benefits will hopefully address the linguistic, technological and connectivity issues encountered by content developers, enabling them to publish highly internationalised and usable digital content more productively and cost-effectively.

6. Conclusion

We started off by distinguishing between physical and digital content in Section 1, and went on to classify digital content as either enterprise or personal in Section 2. In Section 3 we examined why enterprises localise digital content, and the costs involved in doing so. Internationalisation is examined in Section 4, with the three core processes of authoring, enabling and testing explained in detail. Section 5 described how content development guidelines can be implemented into the digital content production process via a Web 2.0 system called the LKR. These guidelines ensure that content is generated for global rather than local audiences, and is less costly to translate and test due to improved translatability (for human translators and machine translation systems) and quality respectively.

The success of the LKR ultimately depends on how well it is embraced by the content developer community. It has the potential to pool the collective knowledge of content developers worldwide, and dramatically increase the quality and consistency of enterprise content being published worldwide. A centralised, online repository of data from subject matter experts in different regions could have enormous benefits for multinational corporations developing digital content for international audiences, and could potentially have an impact on the service fee charged by language vendors in the future.

Despite this promising initial reaction to the concept of the LKR, there are several challenges to overcome. The first challenge is technical; parsers will have to be developed to support the myriad of file formats that content developers currently work with. The second challenge is the usability of the system; it will have to be user-friendly and accessible, allowing users to customise it so that they only access those guidelines which are relevant to their particular projects. Finally, the quality and relevance of the data in Digital Library depends very much on user contributions, interaction and feedback.

References

Aviation Safety Network (2009) 'Accident description', Retrieved 19 June 2009, from http://aviation-safety.net /database/ record. php?id= 19830723-0

Common Sense Advisory (2006) 'Report on global consumer online buying preferences, showing the impact of language, nationality, and brand recognition', Retrieved 19 June 2009, from http://www.commonsenseadvisory.com/news/pr_view.php?pre_id=39

Coyle (2009) 'Facebook gets set for an Irish language lesson', Retrieved 19 June 2009, from Times Online http://www.timesonline.co.uk/tol/news/world/ireland/article5489404.ece

Encyclopaedia Britannica (2009) 'History of computing', Retrieved 19 June 2009, from http://www.britannica.com/EBchecked/topic/130429/computer/216032/Invention-of-the-modern-computer

Ethnologue Languages of the World (2009) 'Ethnologue language name index', Retrieved 09 July 2009, from http://www.ethnologue.com/ethno_docs/distribution.asp?by=size

Europa Languages and Europe (2009) 'Is every document generated by the EU translated into all the official languages?', Retrieved 19 June 2009, from http://europa.eu/languages/en/document/59#5

Facebook (2009) 'Facebook releases site in Spanish; German and French to follow', Retrieved 19 June 2009, from http://www.facebook.com/press/ releases. php?p =16446

Freij, N. (2009) 'Web 2.0 and localization', Retrieved 19 June 2009, from http://blog.globalvis.com/ 2008/04/web-20-and-localization.html

GoToWeb2.0 Web Applications Index (2009) 'Web 2.0 tools and applications', Retrieved 19 June 2009, from http://www.go2web20.net

International Trade Administration (2009) 'Computer software industry 2008', Retrieved 19 June 2009, from http://www.ita.doc.gov/investamerica/ computer_software.asp

International Organization for Standardization (2009) 'Numeric representation of dates and time', Retrieved 19 June 2009, from http://www.iso.org/iso/support /faqs/faqs_widely_used_standards/widely_ used_standards_other/date_and_time_ format.htm

Jobes, K. and Silva, M (2001) Invitation to the Septuagint ISBN 1-84227-061-3, (Paternoster Press, 2001).

LISA (2003) 'LISA industry primer 2003', Retrieved 19 June 2009, from http://www.cit.gu.edu.au /~davidt/cit3611/LISAprimer.pdf

National Aeronautics and Space Administration (2009) 'Mars Climate Orbiter', Retrieved 19 June 2009, from http://solarsystem.nasa.gov/missions/ profile.cfm?MCode=MCO&Display=ReadMore

Netcraft (2009) 'June 2009 web server survey', Retrieved 19 June 2009, from http://news. netcraft.com

O'Reilly Media (2009) 'What Is Web 2.0? Design patterns and business models for the next generation of software', Retrieved 19 June 2009, from http://oreilly.com/web2/archive/what-is-web-20.html

Recording Industry Association of America (2009) 'Consumer purchasing trends', Retrieved 19 June 2009, from http://www.riaa.com/keystatistics.php? content_selector=consumertrends

Software Top 100 (2009) 'The world's largest software companies', Retrieved 19 June 2009, from http://www.softwaretop100.org/list.php?page=1

Tahoe Daily Tribune (2009) 'Another tough year for the music industry', Retrieved 19 June 2009, from http://www.tahoedailytribune.com/article/20090108/ ENTERTAINMENT/901079974/1005/NONE&parentprofile=1056&title=Another%20tough%20year%20for%20music%20industry

Tek Translation (2009) 'Spanish language variations', Retrieved 19 June 2009, from http:// www.tektrans.com/docs/Tek_Educational_Best_Practice_-_Spanish_Variations.pdf

The Guardian (2009) 'Internet data heads for 500bn gigabytes', Retrieved 19 June 2009, from http://www.guardian.co.uk/business/2009/may/18/digital-content-expansion

TidWiT Digital Content Marketplace (2009) 'What is digital content?', Retrieved 19 June 2009, from http://www.tidwit.com/WhatIs.aspx

Trieloff, L. (2007) 'Living in a multiligual world: Internationalization for Web 2.0 applications', Retrieved 19 June 2009, from http:// www.slideshare.net/lars3loff/living-in-a-multiligual-world-internationalization-for-web-20-applications

Supporting Flexibility and Awareness in Localisation Workflows

David Lewis, Stephen Curran, Gavin Doherty, Kevin Feeney, Nikiforos Karamanis,
Saturnino Luz, John McAuley
Centre for Next Generation Localisation
Trinity College Dublin
www.cngl.ie
dave.lewis@cs.tcd.ie; stephen.curran@cs.tcd.ie; gavin.doherty@cs.tcd.ie; kevin.feeney@cs.tcd.ie;
nikiforos.karamanis@cs.tcd.ie; saturnino.luz@cs.tcd.ie; john.mcauley@cs.tcd.ie

Abstract

A key strategy for supporting users in distributed work systems is to help them maintain awareness of the state of the work system and of the work being done by others. At the same time, many knowledge intensive industries are embracing the technologies that have underpinned the Web 2.0 movement to allow open user generation, annotation and modification of content. These technologies can potentially provide a useful platform for supporting awareness and distributed teamwork. However, as distributed content generating activities become more valuable, organisations aim to optimise them, often by modelling and monitoring the workflows involved and augmenting them with software services. Currently, however, these two approaches do not integrate well and there is little system support that integrates the centralised monitoring and management of workflow with the open communications that is characteristic of web-based user content generation. In this paper we examine the use of both techniques in the localisation industry, and based on this analysis we propose a platform that combines the visibility and awareness support of open content generation between users with their involvement in a centrally managed workflow.

Keywords: *localisation workflow crowdsourcing service-oriented meta-data management*

1. Introduction

Situation awareness has long been recognised as critical for supporting effective and resilient performance in complex work systems (Endsley 2000). While the literature has to a large extent concentrated on situations such as process control, command and control etc., supporting situation awareness is an important factor in maintaining the ability of any complex organisation to effectively detect and respond to unforeseen and exceptional situations.

Our analysis focuses on the localisation industry, which translates textual content into different languages so that products and services can be marketed and used in different countries and regions around the world. We start by presenting an abstract description of the localisation workflow whereby such content is translated, based around the view commonly taken by the vendors of workflow products supporting this sector. We then present several examples of the recent trend for crowd-sourcing in localisation, whereby content is distributed for translation to a group of bilingual individuals engaged in a community around the product or service being localised. This is in contrast to the traditional localisation process of outsourcing translation to a professional translation agency.

To provide a more realistic picture of the information exchanged between actors in this domain we present an analysis of the results from a recent field study carried out within the localization industry. Preparatory work included offline study of the available tools, review of background materials on the localization industry (including previous issues of Localisation Focus, the proceedings of the LRC conference and archives of localisation fora such as www.localisationworld.com and www.multilingual.com), and review of the research literature pertaining to localisation including the descriptions of the localisation process by Esselink (2003) and by Wittner and Goldschmitt (2007). A series of 13 semi-structured interviews was then carried out, with interviews typically lasting between 45-minutes to an hour each. The interview subjects were employees of a large company and a multi-language service provider. The interviews were accompanied with observations of the employees using several tools to perform their tasks. The results of this study highlight the interactions that occur both within and external to the formal workflow and its support in the workflow management system. We then compare this analysis to observations made about the communication channels needed in crowdsourced localisation. From this analysis we propose a generic platform for managing workflows with control over the integration of open communication mechanisms related to quality issues. The aim of the platform is to provide greater visibility of the work system, supporting improved awareness, to integrate transient and ad-hoc communications in a more structured manner, supporting knowledge capture, and to support a more flexible and realistic conception of the workflow and work system.

2. Conventional Centralised Localisation Workflow

Figure 1 depicts a generalized localization workflow. This process is based on one of the author's experience (Stephen Curran) working as a software engineer in localisation. The chain of activities in the workflow is as follows:

- **Extraction:** The process of extracting translatable text from the source documents. Documents coming from desktop publishing packages often contain a lot of structural information that is not needed for the translation process. Textual content is separated from structural content.
- **Segmentation:** The process of dividing up the source document into translatable units of text. Normally these are sentences but they don't have to be. They could for instance be paragraph headings or diagram captions.
- **Creation of Project TM:** The segmented documents are analysed against the central translation memory to produce a project translation memory. The project translation memory contains all relevant translations from the central translation memory.
- **Pre-translation:** The target of every exact match in the project translation memory is inserted into the translation placeholder of the corresponding segment in the document. This is an optional step in the workflow. Sometimes it is preferable to have the translator manually insert the exact match from the project translation memory into the document themselves so that the translation unit is getting a certain amount of review.
- **Machine Translation (MT):** Sometimes a machine translation system is used to generate translations for segments in the documents that do not have any match in the translation memory.
- **Generation of Translation Kit:** All files needed to perform the translation are zipped up and sent to the translator. This includes the documents to translate and the project translation memory. It may also include a glossary containing any relevant terms and their translations and any reference material required to give context for the translation.
- **Manual Translation:** The translation kit is downloaded and unzipped by the translator. The translator opens the documents, translation memory and glossary in their translation environment and iterates through and provides a translation for each document segment. When the translator opens a segment in the environment any match in the TM or glossary is presented to the translator for insertion into the target segment along with a notification of the match value.
- **Review and Editing:** The translated documents go through a cycle of review and editing. This includes both linguistic review and functional testing.
- **Translation Memory Update:** Once the documents have been signed-off from review the translation memory is updated with the translations from the documents. The updating process includes inserting any new translations and updating any previous translations.
- **Creation of Target Documents:** The final target version of the documents are created. The translated segments are combined with the document structural information to produce the final version of the documents.

3. Crowd-sourced Localisation

The localisation industry is increasingly turning to crowd-sourcing to address the scalability problem of current processes. In localisation crowd-sourcing, the translation job traditionally done by a professional translator is done in a more informal fashion by a group of volunteers. These volunteers are usually engaged in a community around the product being translated. Some notable organisations have adopted in varying degrees the crowd-sourcing approach to localisation :

Facebook

The social networking site Facebook crowdsources the translation of their user interface (Facebook Translations). Users, through a Facebook application, can submit translations for strings in the user interface. A translation memory is incrementally built and made use of by the community. Quality control is achieved through the community commenting and rating each other's translations.

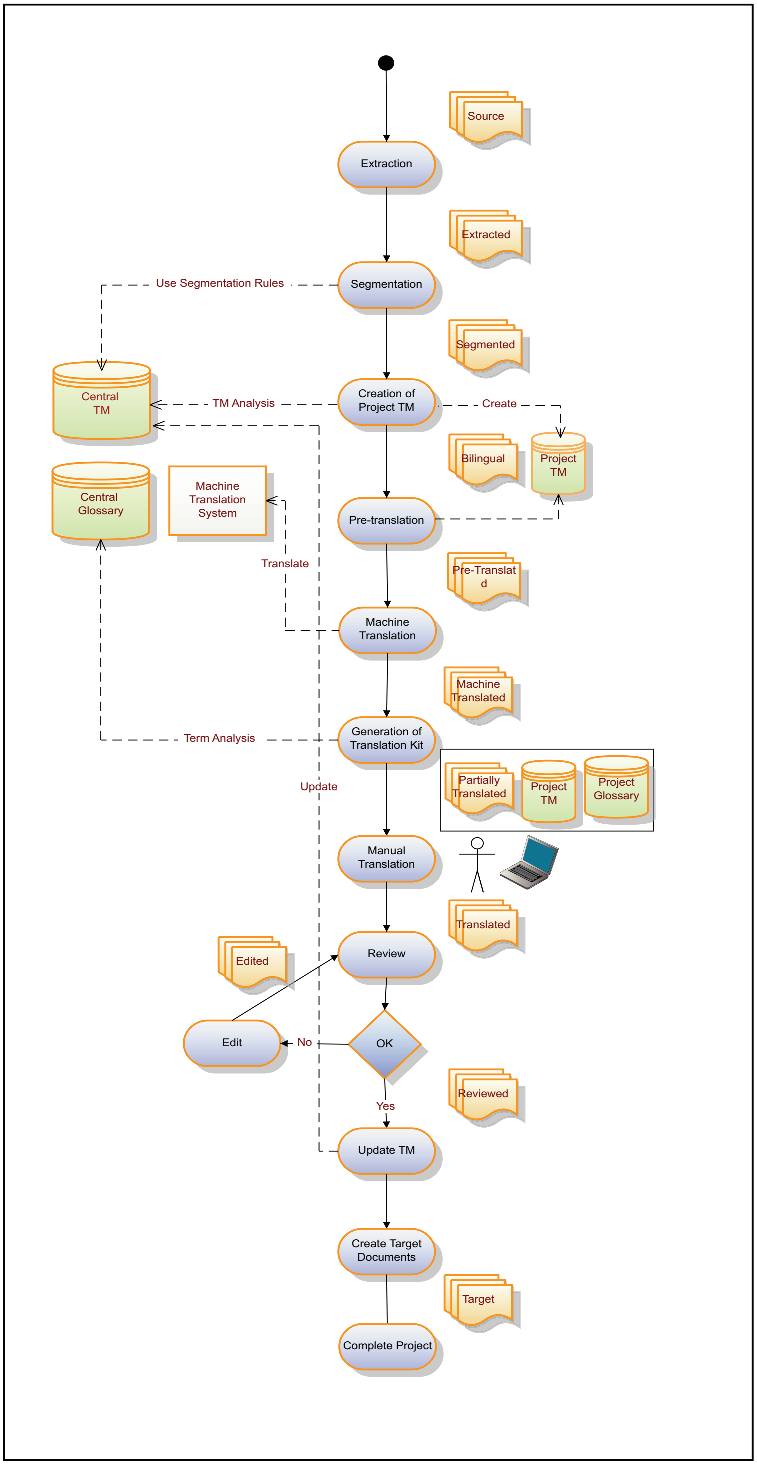


Figure 1: Generalized Localization Workflow

We use this model as the basis for conducting a study of the actual interactions that are part of the current practice of the interviewed practitioners. These interactions include informal communications as well as those that are part of formal, business processes supported by workflow tools.

For each target language, the community translation process goes through the following steps:

- The core terminology is translated by the community to build a bilingual glossary.
- The user interface strings are translated by the community making use of the glossary.
- The contributed translations are rated by the community and winning translations are determined.
- The translations pass through an internal review stage before being approved.

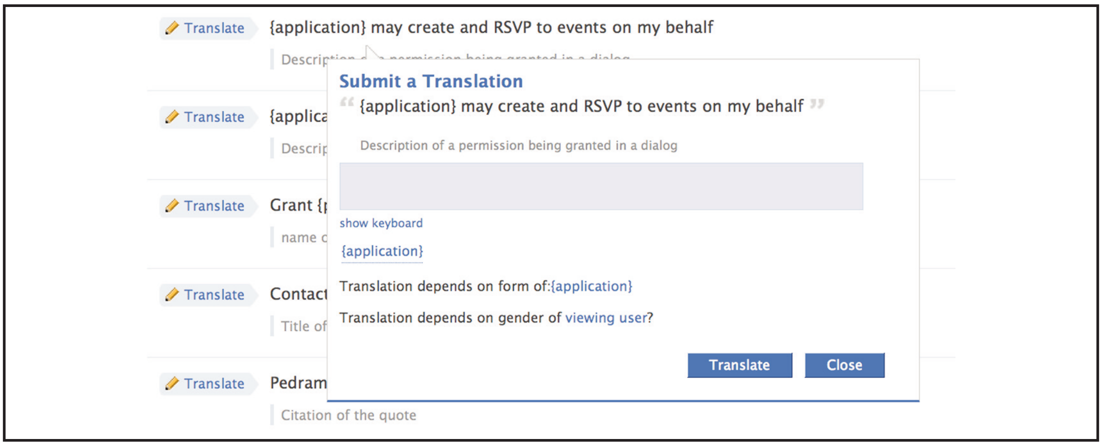
Microsoft

Microsoft has a forum on its MSDN website that allows the community to contribute and rate translations of terminology in Microsoft products (Microsoft's Terminology Community Forum). Better translation of product-specific terminology can often be forthcoming from the product user community than from a professional translator who does not use the product. Whereas Facebook crowdsources the translation of sentences, Microsoft crowdsources only the translation of core terms. The community are not involved in the translation process other than providing suggested translations for these core terms.

TED Translations

TED is a non-profit organisation that runs conferences on topics in technology, entertainment and design. Recently, in an attempt to reach out to a wider audience, they have published their conference presentations in languages other than English. They have used a crowd-sourcing approach to translate the presentations to these languages (TED Translations).

TED, like Facebook, get the community to translate sentences rather than just terminology. However, unlike Facebook, the translator tends to work on the entire text rather than on small segments. There is no concept of on-going rating or feedback of other parallel translation taking place within the text. The translator tends to work independently on the entire text.



Translate {application} may create and RSVP to events on my behalf

Description of a permission being granted in a dialog

Translate {applica

Descrip

Translate Grant {p

name c

Translate Contact

Title of

Translate Pedram

Citation of the quote

Submit a Translation

“ {application} may create and RSVP to events on my behalf ”

Description of a permission being granted in a dialog

show keyboard

{application}

Translation depends on form of:{application}

Translation depends on gender of viewing user?

Figure 2: Facebook translation interface

	Community involvement	Interaction in the community	Resources that the community interacts with
Facebook	High. Full involvement in the translation of glossary and the translation and review of product.	High. The community discuss and rate translation.	Translation memory and glossary.
Microsoft	Low. Input into the translation of glossary. No involvement in the translation and review process.	High. The community discuss and rate translation.	Just glossary.
TED	High. The community translates all of transcript.	Low. A transcript normally translated by one individual.	No resources used. Translation memory and glossary not maintained.

Table 1: Analysis of Translation Crowd-sourcing Systems

Unlike Facebook, no translation memory or termbase is maintained. The language used in the talks is not controlled and the topics of the talks are varied, so its not clear that there is much benefit in using a translation memory or terminology manager.

Summary of approaches

Table 1 summarises these approaches and their relative involvement with the communities concerned. A key distinction from conventional localisation workflows is the emphasis on supporting peer interaction within the community. Supporting the development of a sense of community by enabling different translators to communicate freely contributes to a sense of shared endeavour that serves to motivate translators in the absence of direct financial reward for their efforts.

From a systems point of view, crowd-sourcing platforms place an emphasis on promoting communication between those involved. This contrasts with workflow management systems which only concern themselves with task related control and data flow between participants, and therefore tend to ignore communication and information sharing that occurs outside of the workflow.

However, as has often been observed, workers involved in workflow often encounter problems that are not fully modelled in the prescribed procedure to which they are expected to work. They therefore often seek solutions by using informal communication with other workers. In a shared office environment such communication can take place readily, but in workflows where human knowledge based activities within a process are undertaken in different organisations and by geographically distributed staff, such informal communication may be less easy to initiate as workers are less likely to be personally acquainted. The ability of web based communication technologies to support and encourage open communication between unacquainted crowd-sourcing workers may therefore offer some benefits to commercial workflow systems. To understand better what potential channels of communication could be beneficial to existing commercial localisation workflows, we look at the current working practices and problems revealed in the study. We examine those aspects that are not addressed in current localisation workflows and are therefore not implemented in the systems that support them.

4. Role Interactions in Localisation Workflow

Across the centralised, commercial localisation workflow and crowd-sourced localisation the following abstract roles can be identified:

- Content author: produces the source text
- Terminology Manager: manages a consistent set of terms and expressions for a project in the source language.
- Linguist: a language specific specialist responsible for maintaining translations of terms, translation manuals and the guidelines used in assessing quality.
- Project Manager: overviews the translation process on behalf of the clients. In a commercial setting this is sometime the activity of third party LSPs, while in a crowd-sourced setting this may take the form of an online community moderation role.
- Post-editors/Translators: manually translate source text to target language text. Post editing involves reviewing existing translations, whether produced by human or automated translators and involves selecting from alternatives, modifying a translation or providing a new one.

The following table illustrates the points of interaction between these identified roles. The interactions marked "I" are informal ones that are poorly supported in current localisation processes while the interactions marked "W" are those that are already supported in existing localisation workflow systems. Those marked "N" are ones that arose through discussion as being of potential value but which are not currently common practice. What is clear from this analysis is that current workflow systems and processes focus on communication that follows the workflow process from one role to the next. What is not well supported is upstream communication from the roles operating at later portions of the workflow to those involved at the earlier parts. Although such interactions may help improve the overall process, they lie outside of the main flow of communication. Since the workflow model is seen as the primary route to value generation and therefore the basis for contractual arrangements, this means little relative value is attached to these communications.

To/From	**Terminology Manager**	**Content Author**	**Linguist**	**Project Manager**	**Translator / Posteditors**
Terminology Manager	Share term bases and techniques for achieving high compliance to controlled language guidelines (N)	Detail problems encountered with applying the term base and controlled language guidelines (I)	Relate language-specific problems in translating specific terms from term base (I); propose changes to controlled language guidelines to improve efficiency of translation to a specific language (I)	Relate problems with conformance to controlled language and missing terminology (via linguist) (I)	Relate problems with conformance to controlled language and missing terminology (via project manager) (I)
Content Author	Term-base and controlled language guidelines (W)	Share notes about complying with controlled language guidelines and appropriate terminology	Specify the job target level for controlled language compliance (I)	Relate translation problems with provided content and its context (I)	Errors in source content and missing contextual information (I)
Linguist	Response to stated translation problems with specific terms (I)	not applicable	Share problems in translating term base to different languages (I)	Quote for job (W); Relate problems with using terminology translations (I); Problems with use of translation guidelines (I)	Suggest different terminology translations (N)
Project Manager	Term base and its context (W)	Content and its context (W)	Translation dictionary and guidelines (W)	Share problems with translating jobs into parallel languages, handling specific content, performance of specific TM, MT and translators (I)	Progress in translation job (W); Problems in translating content, erroneous content, terminology or term translation (I)
Translator	Term base and its context (via project manager) (W)	Content and its context (via project manager) (W)	Translation dictionary and guidelines (via project manager) (W); Responses to translation problems with terminology (I)	Job allocation and quality targets (W)	Share problems with use of terminology translations and lack of terminology definition/translation; queries to more experienced translators; Feedback on quality of TM and MT translations (I)

Table 2 : Summary of Role Interaction in Localisation Workflow

5. System Support for Localisation Workflow Awareness

It seems clear that the future of localisation will involve some element of crowd-sourcing. However, to reach its full potential of crowd-sourcing to optimally complement commercial localisation activities, such next-generation localisation will need new integrated platforms that integrate crowd-sourcing technologies and workflow management technologies together seamlessly. The study shows that current commercial workflow could benefit from improved communication that goes beyond that dictated by the major value flow of the workflows. However, the very open communication characteristics of the crowd-sourcing environment are not always appropriate for workflows operating within the constraints of commercial contracts. A client may make use of several Localisation Service Providers (LSPs), perhaps with different financial arrangements, and may therefore be sensitive to free communication between them. Equally, LSPs compete for clients and may not be willing to expose the less formal, but valuable interactions that have built up with a particular client.

Furthermore, any common platform for supporting both commercial and crowdsourced localisation would need to support the evolution and migration of workflows between the two for both companies and individuals, as many will be involved in both, to varying degrees at different times.

The challenge in developing a common model for integrating commercial and crowdsourced localisation is to support a variety of levels of control over the interactions that can occur with a localisation process, ranging from tight central control to loose, highly devolved control.

The Centre for Next Generation Localisation (CNGL) is an integrated research project bringing together academic researchers and industrial partners to construct a framework to support the next generation of localisation systems that precisely encompasses both commercial and crowd-sourcing localisation in order to support a growing variety of new business models that can deploy both. It remains unclear whether the integration of features designed specifically for managing human tasks directly into the workflow management system is desirable as such an approach increases the complexity of the workflow specification, yet will significantly constrain the expressiveness that can be applied to human task management.

Below, we outline an integrated software architecture that is being assembled within the CNGL project to support standards based web service integration and workflow execution with flexibility in how the human communities involved in such workflows can communicate with each other.

Central Meta-data Repository

It was identified earlier in this paper that many of the transient communications in current localisation workflows are being lost because the workflow technology does not support them. In this section we introduce a potential solution to the problem.

On the internet there has been a move towards adding structured information to content so that it can be automatically reasoned over. With the semantic web initiative content publishers are being encouraged to attach additional meta-data to the resources that they publish. The Resource Description Framework (RDF) is a meta-data data model that is being used to support this initiative (RDF 2004). RDF allows for the representation of meta-data in the form of 3-place relations, subject, predicate and object. The data model can therefore support any arbitrary data schema. This is necessary as it is not known a priori the range of meta-data that content publishers will want to express.

We propose using an RDF repository to store communications in and across organisational boundaries. A flexible meta-data schema is necessary since it is also not known the range of communication that might need to be represented in a localisation workflow. More precisely, each resource in the localisation workflow : TMs, glossaries, controlled-language rules etc. would have a set of communications associated with it that are stored in the repository. Storing the information centrally means that it can be easily aggregated and reasoned over. Since potentially many organisations could be involved in a localisation workflow and could contribute to this store, we propose using our Community Based Policy Management technology (see below) as a means of managing this store in a decentralised way.

Community Based Policy Management

Given the cross-organisational nature of localisation workflow, the communication meta-data repository would consist of communications within and between organisations with no one organisation owning all the communications. Therefore, each organisation should be able to manage the communications that it owns and control which other individuals and organisations should have access to it. Our Community Based Policy Management (CBPM) technology (Feeney et al 2004) can be used to allow the repository to be managed in such a decentralised fashion.

The CBPM is a policy based management system which allows for the sub-division of organisations into smaller groups. Each of these groups has its own set of policies applied to it, meaning that each group has a certain set of rights over the resources owned by the organisation. CBPM also supports the delegation of management rights to various sub-groups or federated groups, meaning that management can be decentralised.

Community Management Framework

We have taken the Drupal Content Management System (CMS) (Drupal 2009) and integrated it with CBPM. Drupal is a web content framework with a pluggable architecture and a collection of add-on modules contributed by a development community. These modules include such things as forums, messaging, blogs and other social networking and communication technologies.

The integration of Drupal and CBPM allows online communities to control the distribution of management authority over content in the CMS across the community. Since Drupal comes with these communication technologies, the Community Management Framework (CMF) can be used in a localisation crowd-sourcing scenario to help manage communication between volunteer translators in a fine grained manner. This allows community management decision makers (who may vary from professional community moderators to a democratic function of the whole community) to balance the benefits of completely open communication with those of more restricted, team based communication and to move easily from different models as the focus and activity level of the crowd-sourcing community shifts over time.

Business Process Execution Language and Human Tasks

A Service Oriented Architecture approach has been taken in development of the CNGL project demonstrators. Business Process Execution Language (BPEL) is used as a platform for creating and executing localisation processes. BPEL automates business processes through the orchestration of web services. Linguistic processing software components, performing functions such as Machine Translation or Text Analytics, are packaged as web services and used by BPEL processes. BPEL is good for task automation but the central standard does not support human tasks. In localisation processes, some tasks are manual: professional translation/post-editing, crowdsourced translation. We need a way to support the inclusion of such tasks.

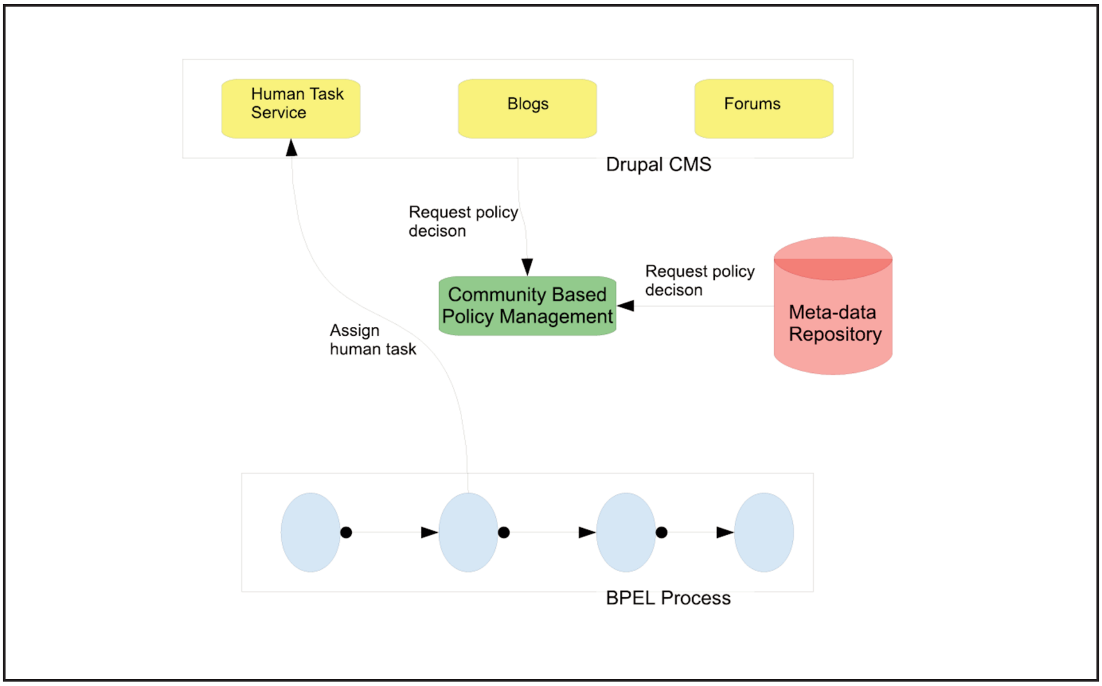


Figure 3: Integrated architecture for Flexible Interactions between Localisation Workflow Actors

Currently, human tasks are integrated into BPEL-based workflows through a human task web service that is included in the workflow as a partner service. The central BPEL process sends the task to this service and the task is completed by people interacting with the service. The web service implements such functionality as task assignment, task workflow and task lists. BPEL execution engines normally come with such a human task web service.

There is also an OASIS standard, BPEL4People, that aims to include support for human tasks directly in BPEL. The advantage of this is that humans and task assignments can be modelled directly in BPEL. However, there is very little support for this standard in BPEL execution engines.

We aim to create a human task web service and integrate it into the CMF so that through the application of CBPM rules, the tasks will be routed to the appropriate individuals within the organisation. This may be useful in the localisation crowd-sourcing scenario where appropriate translators can be selected based on policies in CBPM, taking into account such criteria as source and target language, the domain of translation, and the reputation of the translator.

Architecture for System Support for Flexible Interaction between Localisation Workflows Actors

Hence our integration framework includes a custom service interface designed to support the expression of a wide range of the management requirements that organisations typically wish to apply to human tasks. This human task service is distinguished from other implementations in that it provides native support for task delegation and decomposition within groups, rather than requiring the workflow designer to specify the intimate details of how each task should be performed and monitored. It is based on the CBPM, and allows tasks to be allocated to groups or communities in addition to individuals. These groups can then use the CBPM system to further break down tasks, allocate them to individuals and gain fine-grained control over the process. This ability to delegate portions of the workflow is crucial in supporting large, complex workflows that span organisations without requiring the workflow designer to have a complete understanding of every detail of the process.

6. Discussion and Further Work

In this paper we have discussed a number of issues (identified through fieldwork) relating to traditional workflow management systems, including the opaqueness and coarse granularity of the work, the lack of support for the continuous and out-of-band communications which are a feature of effective teamwork, as well as departures from the formal workflow needed in order to deal with changing circumstances or exceptional cases. We have analysed these issues with respect to the localisation process, by examining a number of role interactions, and the degree to which they are supported in the workflow. These interactions are not just between individuals in a role, but between individuals and groups, between groups and between different subsets of groups and the rest - especially when groups (or individuals with them) play more than one role.

We have examined the recent trend towards distributed content generation and management, which presents an opportunity to leverage the same technical infrastructure across a range of systems, ranging from centrally controlled to fully distributed. Localisation provides an excellent example, as there is an opportunity to achieve integration between crowdsourced and enterprise localisation technologies.

The solution proposed here is based on adding metadata and additional links to existing artefacts beyond the major transactions of the workflow. Within the context of localisation, a range of artefacts (terminology, TM entries, source text) that are exploited across the process can provide a vehicle and a set of interface concepts for achieving this integration. This approach will help to make work, effort and resources more visible, which will increase awareness throughout the workflow, and will also allow the many exceptions to be dealt within the formal structures of the system. We are in the process of prototyping such a system and will aim to trial it in a real localisation scenario in the future. While dynamic communication features are becoming a common feature within workflow systems, these should be integrated in a way that facilitates knowledge capture, so that they become a resource for the rest of the team.

Acknowledgements

This material is based upon works supported by the Science Foundation Ireland under the Centre for Next Generation Localisation (www.cngl.ie).

References

Allee, V. (2002) The Future of Knowledge: Increasing Prosperity through Value Networks, Burlington: Butterworth-Heinemann

Drupal Content Management System (2001) [online], available: http://www.drupal.org [accessed 6th Feb 2009].

Endsley, M.R. (2000) 'Theoretical Underpinnings of Situation Awareness: A Critical Review', in Endsley M.R and Garland D.J. (eds.) Situation awareness: analysis and measurement, New Jersey: Lawrence Erlbaum Associantes, Inc., 3-33

Esselink B. (2003) 'Localisation and Translation', in Sommers H., ed., Computers and Translation, Amsterdam: John Benjamins, 67-87.

Facebook Translations Application (2008) [online], available: http://www.facebook.com/apps/application.php?id=4329892722 [accessed 25 Jan 2010]

Feeney, K., Lewis, D., Wade, V. (2004) 'Policy Based Management for Internet Communities', in Proc. of IEEE 5th Int'l Workshop on Policies for Distributed Systems and Networks, New York, 7-9 June, IEEE Computer Society Press, 23-34.

Microsoft Terminology Community Forum (2008) [online], available: http://www.microsoft.com/language/mtcf/mtcf_default.aspx [accessed 25 Jan 2010]

OASIS, WS-BPEL Extension for People (BPEL4People) (2009) [online], available: http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=bpel4people [accessed 25 Jan 2010]

Resource Description Framework (RDF) (2004) [online], available: http://www.w3.org/RDF/ [accessed 24th July 2009]

TED Translations (2009) [online], available : http://www.ted.com/translate [accessed 25 Jan 2010]

Web Services Business Process Execution Language (BPEL) Version 2.0 [online], (2007), available: http://docs.oasis-open.org/wsbpel/2.0/OS/wsbpel-v2.0-0S.html [accessed 6 Feb 2009]

Wittner J. and Goldschmidt D. (2007) 'Technical Challenges and Localisation Tools', Multilingual #91, Localisation Guide: Getting Started.

Applying Digital Content Management to Support Localisation

Alexander O'Connor[1], Séamus Lawless[1], Dong Zhou1, Gareth J. F. Jones[2], Vincent Wade[1]
[1] Centre for Next Generation Localisation
Knowledge & Data Engineering Group
School of Computer Science & Statistics
Trinity College Dublin
Dublin 2, Ireland
[2] Centre for Next Generation Localisation
Dublin City University
Dublin 9, Ireland
www.cngl.ie
Alex.OConnor@cs.tcd.ie, Seamus.Lawless@cs.tcd.ie, Dong.Zhou@cs.tcd.ie, Vincent.Wade@cs.tcd.ie
Gareth.Jones@compuing.dcu.ie

Abstract

The retrieval and presentation of digital content such as that on the World Wide Web (WWW) is a substantial area of research. While recent years have seen huge expansion in the size of web-based archives that can be searched efficiently by commercial search engines, the presentation of potentially relevant content is still limited to ranked document lists represented by simple text snippets or image keyframe surrogates. There is expanding interest in techniques to personalise the presentation of content to improve the richness and effectiveness of the user experience. One of the most significant challenges to achieving this is the increasingly multilingual nature of this data, and the need to provide suitably localised responses to users based on this content. The Digital Content Management (DCM) track of the Centre for Next Generation Localisation (CNGL) is seeking to develop technologies to support advanced personalised access and presentation of information by combining elements from the existing research areas of Adaptive Hypermedia and Information Retrieval. The combination of these technologies is intended to produce significant improvements in the way users access information. We review key features of these technologies and introduce early ideas for how these technologies can support localisation and localised content before concluding with some impressions of future directions in DCM.

Keywords: *Digital Content Management, Information Retrieval, Adaptive Hypermedia, Content Analysis, Open-Corpus Content, Multilingual Technologies*

1. Introduction

Digital Content Management (DCM) is concerned with the creation, transformation, storage, retrieval and presentation of information in digital form. At present, the most publicly visible resource available to DCM applications is the World Wide Web (WWW). The current approach to content management for web applications is very limited by the assumption that content is largely static and by providing access via search engines which broadly assume static file collections held individually on specific servers. However, this is rapidly becoming an outdated model of the way that most information exists on the web. Static file structures are giving way to web-based content-management systems, which compose responses dynamically using content stored in databases. This content can be presented to the user in different ways depending on style, accessibility or security preferences. The web itself is becoming a collection of highly-diverse content management mechanisms. This is creating substantial challenges to the satisfaction of user information needs because this heterogeneity of data sources introduces complex obstacles to computational methods for managing content. In addition, with nearly 500 billion gigabytes of information being stored worldwide (Wray 2009), the need to be able to find and index specific information has become a massive global challenge. In order to be able to better support users with complex information needs, it is also necessary to develop new ways of responding to users that go beyond the conventional ranked listing of documents.

Our belief is that an effective way to address these challenges can draw on two principal areas of research: Information Retrieval (IR) and Adaptive Hypermedia (AH). Research in IR underpins existing search engines such as Google, and enables efficient search for relevant documents among the billions of items currently available on the web. A particular challenge in the selection and presentation of this content is the increasingly multilingual nature of digital content. Effective DCM systems need not only to find and present content, but they need to do this in a multilingual environment with the output ultimately in a form that can be reliably and comfortably consumed by the user. Search and presentation already presents challenges; extending this to more personalised formats is considerably more demanding. AH research meanwhile focuses on the view that the power of digital content is its malleability. AH technology takes as its goal the creation of highly tailored, rich media presentations designed for the specific needs of the user. AH technology has its roots in eLearning systems, which teach complex concepts to students using rich media experiences. The main limitation of current AH techniques is that they have to-date focused on small, carefully controlled content sources, making them unsuitable for highly heterogeneous data sets such as the WWW. Further, as digital technologies proliferate, there is a compelling need to address the issue of documents authored or stored in different natural languages.

Localisation, based on the manual or machine translation of content, is thus a major concern and opportunity for DCM. Localisation is a mature domain with substantial industrial experience in many issues associated with managing corpora of content in different languages designed to serve different groups. It is impossible to attempt to address the challenge of effective global DCM without also addressing the language and localisation of content selection and presentation. DCM technologies thus aim to provide new functionalities for addressing emerging opportunities and challenges of localisation for dynamic multilingual content.

Work in DCM within the CNGL is seeking to integrate IR and AH technologies with language translation and input derived from existing experience with localisation. The goal of this integration is to develop novel and effective technologies for personalised responses to user information needs; taking data from open, multilingual heterogeneous data sources. This paper introduces some of the key background technologies which are being investigated within the DCM track of the CNGL. The paper begins with background reviews of AH and IR, and of some of the specific techniques being used to make the wide variety of content available on the WWW more accessible. With these technologies outlined, the paper then describes some potential applications for different combinations of these technologies in the specific area of localisation. The paper concludes with some remarks on potential future directions in DCM research.

2. Adaptive Hypermedia

Conventional static web content management systems present the same responses to all users regardless of their preferences or other personal factors. However, these classic 'one size fits all' content delivery systems are simply not powerful enough, particularly as the WWW becomes increasingly dynamic and multilingual. Web delivery systems and hypermedia systems are increasingly attempting to customise content so that it is relevant to the user or the context of use. This can be achieved by, for example, changing the presentation of the content for different screen formats, or by allowing users to alter its layout manually. These technologies separate some elements of content from presentation, for example by using stylesheets which can take account of personal or localised needs. However, the majority of these systems do not go far enough towards meeting individual needs.

We believe that AH technologies can make an important contribution to extending the limitations of current web-based content management systems. Such new systems can make it possible to deliver "personalised" views of a hypermedia document space without requiring programming from the content author. This is achieved by building a model of the goals, preferences and knowledge of the individual user, and using this model to dynamically compose responses tailored to the individual user. For example, a large component of the value of digital content in elearning is in the targeted delivery of that content to the right user in the right form. This is typically achieved using three specific component models: a user model, a content model and a domain model (Conlan 2004).

User models can be initialised by explicitly illiciting information from the user using a questionnaire or through the use of stereotypical user models. User models can also be evolved automatically through adaptation by simply observing the browsing behaviour of the user. In this way, the user model can continually adapt as the user works with the system and their preferences and knowledge develop.

At present, hypermedia systems are generally restricted to the use of "closed" content collections. It is assumed that the content to be used by these systems is authored so that the collection consists of pieces of content each of which covers some number of related concepts within a subject. These content pieces are typically annotated with highly structured metadata describing various features of the content. There are a variety of international standards for such document description. Some standards, such as Dublin Core (Dublin Core), support the creation of metadata to describe a document in a domain agnostic fashion. Some are more domain specific, such as Learning Object Metadata (LOM) in the eLearning domain. This metadata is generally added manually, meaning that the cost of producing content for use in AH systems is often very high. More recent systems focus on automatic, or semi-automatic generation of metadata as part of the content authoring process. Others focus on generating the metadata based on the context within which the content was originally developed. A third approach, explained below, focuses on inspection of content chunks to facilitate the generation of the metadata. However, manual annotation (metdata tagging) is still quite common due to the importance of accurate high quality metadata in AH systems. Because of the expense of authoring such content, one of the goals of research for AH systems is to maximise the exploitation and reuse of content in order to recoup the return on investment of content creation.

The third element of most AH systems, the domain model, contains a conceptual description of the subject area(s) of interest and a specification of the relationships between these concepts. By dynamically combining the domain model, user model and content model, AH systems can generate personal navigations of adaptively retrieved relevant content.

Having a detailed knowledge of the content when the system is designed is obviously rather different to the situation of the designer of an IR system where the designer often has very little knowledge of the features of the content which is to be indexed or searched. The challenges resulting from these contrasting approaches are addressed in more detail in section 4 of this paper. A good introductory review of current AH technology is contained in (De Bra 1999) and (Brusilovsky 1996).

AH systems use various models to generate a navigation through dynamically (adaptively) retrieved content. In addition to the user model, content model and domain model, more recent AH systems have also begun to use other models e.g. models describing the context within which the user is seeking information and a model of the device upon which the retrieved information is to be viewed (Conlan 2004).

2.1 Adaptive Hypermedia Functions

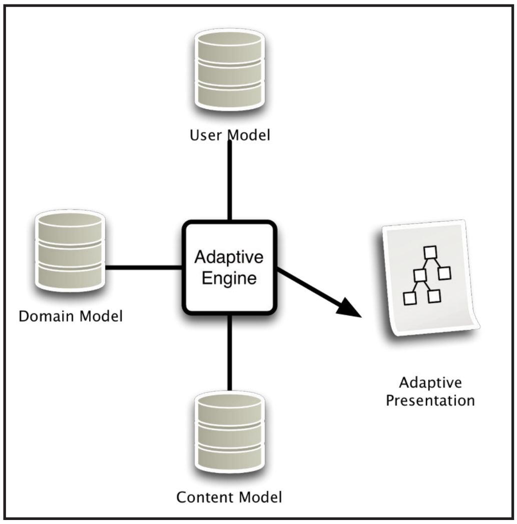


Figure 1: An Adaptive Hypermedia Framework, showing the influence of three models contributing to the creation of an adaptive presentation

An AH system can be thought of as supporting three functions:

- While the user is interacting with the system all, or selected, user actions are registered. Based on these actions, previous behaviour and other user-supplied or automatically-gathered information, the system builds a model of the user's knowledge about each domain model concept. The system seeks to model how much knowledge the user has about the concept and what information they have read about it, as well as other attributes about the user which can be used to tailor the adaptation process.
- The adaptive system reconciles the user model to classify all nodes (content pieces) into one of several group's depending on the user's current

knowledge, interests and goals. The system manipulates links within nodes (and link destinations) to guide users towards appropriate, interesting and relevant information. This is called adaptive navigation in (Brusilovsky 1996).

- In order to deliver the content of a page at an appropriate level of difficulty or detail the system can conditionally show, hide, highlight or dim page fragments. This process is referred to in (Brusilovsky 1996) as adaptive presentation. This is an important component in AH systems and is outlined in the next section.

2.2 Adaptive Presentation

Selecting content for presentation to the user is only part of the functionality of AH systems. In addition, they typically include elements of adaptive presentation. The selection of these pieces can have different consequences for the final presentation depending on the range of mechanisms employed to create the presentation. Adaptive presentation can enable:

- the provision of fundamental, additional or comparative explanations: For example the AH system can add important background knowledge for novice users. This can be achieved in two primary ways; either by including additional content, or by allowing for 'stretchtext'. In the first case, the system can attempt to predict which additional information might be needed by the user based on their user model, and create a presentation which includes additional content pieces directly in the text (De Bra 1997). In the second case, the user can click on a particular term or element of the presentation, and the system will adaptively retrieve the appropriate content and present it separately.
- the provision of explanation variants: Depending on the user model, a variety of elements can be adapted: the level of difficulty, the links to related concepts, the length of the presentation, or the media type (text, images, audio, video). This can be done within a page or through guidance towards different pages in a process referred to as adaptive navigation support.
- the re-ordering of information: The user model can be used to vary the order in which information is presented to the user, similar to ranking in information retrieval. This can be used, for example, to create shorter or longer presentations, or to create presentations which are easier to browse by having summary information appear first.

AH is not just dependent on the existing hyperlinks within a document (or content piece). Adaptive link insertion allows for new, dynamically generated paths through the content space to be generated. This provides the appearance of new aggregrations of hyperlinked documents which are formed just-in-time for a particular user. This allows an AH sytem to, for example, annotate different links with a 'Traffic Light' metaphor, where the system can help the user make navigation choices that best suit their knowledge and preferences. This helps to keep the progress of a user through the content smooth and consistent. Several examples of adaptive navigation support can be seen in (Brusilovsky 2004).

3. Multilingual Information Retrieval

The standard aim of IR is to satisfy a user's information need. IR systems attempt to fulfil this objective by returning a list of potentially relevant documents to the user. In order to satisfy the information need the user needs to extract the information from the documents, typically by reading them. Depending on the information need and the structure and contents of the documents, this can be a very efficient or very inefficient process. An example of how this process can be very inefficient is if the document contains large amounts of information which is not of interest to the user, either because it is off topic or because the user is already familiar with most of it. If the user is seeking small new details buried deep within the content, this becomes a very labour intensive process. We believe that AH methods have scope to begin to address these weaknesses of current IR methods.

Algorithms for IR are typically based on statistical techniques which count the frequency or rarity of words in document collections. The most popular standard measure for this approach is referred to as the tf.idf weighting scheme (Term Frequency - Inverse Document Frequency) (Salton and Buckley 1988). Tf.idf measures the importance of a particular term in a document when considered relative to that term's importance in the overall corpus. This is achieved by combining the term frequency within an individual document with its distribution in the collection as a whole. A simple document ranking can be produced by summing weights for terms (words) occurring in both the user's query and each document. Statistical techniques such as these also feature in probabilistic models, such as BM25 (Spärck Jones et al. 2000a) (Spärck Jones et al. 2000b), which shows increased resistance to noise in identifying particular documents in a collection.

In addition to ranking documents based on the match between a user's query and the document content, for hyperlinked document structures, such as the web, algorithms such as HITS (Kleinberg 1999) and PageRank (Brin and Page 1998) take advantage of the network structure to determine document significance dependent on their place within the network, but independent of their content. Web search engines thus provide an overall ranking of documents in response to a user search requests by combining content matching scores with a network-based measure of the document's likely importance to the user.

Multilingual Information Retrieval (MLIR) moves IR beyond the situation where the query and documents are expressed in a single language, to environments where documents may be a range of languages and queries can be performed in any of these languages. Within MLIR much attention has been focussed on the simpler problem of cross-language (or bilingual) information retrieval (CLIR) where search topics or requests in one language are used to retrieve documents in one other language.

Supporting MLIR requires two main features: adapting IR methods to each document language and developing strategies for translating between query and document languages to cross the language barrier. While early work in MLIR concentrated on text documents from published news sources, more recent work has extended this to explore various multimedia IR data sources including annotated photographic and medical images, spoken data sources, and multilingual web documents (CLEF).

In the case of CLIR there is the inescapable additional issue that while a retrieved document may be relevant to the information need, the user may not have sufficient knowledge of the document language to be able to identify it as such, and to extract the information they are seeking from it. Machine Translation (MT) or content gisting in context, based on bilingual machine-readable dictionaries, has been investigated as a means of accessing particular information or at least determining whether a document is relevant (Oard and Resnik 1999)(He et al. 2003)

MLIR is potentially a key technology in supporting the localisation of dynamic, or rapidly published content. MLIR can aid localisation by making content available across languages in response to user search queries and potentially by providing translation between queries and documents. It can also support the presentation of retrieved content in a culturally sensitive way. The next section summaries some of the key research in the area of CLIR.

3.1 Cross-Language Information Retrieval

In CLIR there is a linguistic mismatch between the queries that are submitted and the documents that are retrieved. To resolve this mismatch CLIR systems incorporate some facility for content translation to bridge this language barrier, an obvious requirement if query representations and document representations are to be meaningfully compared. The performance of a CLIR system is heavily reliant upon the success of this translation process, and therefore the tools and techniques used for automatic translation have formed much of the focus of the CLIR research community.

One of the main questions that arises when addressing a CLIR task is whether to translate the queries to the language of the documents, or the documents to the language of the queries. There are pros and cons with each approach. For example, translation of documents may be more reliable since there is extensive contextual information available. However, query translation does not increase the content storage overhead. Extensive experiments have been carried out comparing document translation and query translation, and the combination of both (McCarley 1999)(Oard 1998)(Oard and Hackett 1997). While document translation and combination methods can outperform query translation, these results have generally been set aside due to the prohibitive effort required to translate the complete document collection into the query languages to be supported and the storage of the index data associated with them. Thus, the overwhelming majority of CLIR systems today operate via query translation.

Typically, three types of resources are widely adopted for translation in CLIR: bilingual wordlists (or machine readable dictionaries) (Adriani 2000)(Gao 2002)(Liu 2005)(Zhou 2008), parallel texts (Chen and Nie 2000)(Nie 1999), and machine translation (MT) systems (Kwok and Dinstl 2007)(Wu 2007). Query translation has most often been effected using machine readable bilingual dictionaries. Unfortunately, bilingual dictionaries have an inherent tendency towards ambiguity. This problem stems from the choice of possible translations. A typical bilingual dictionary will provide a set of alternative translations for each term within any given query. Choosing the correct translation of each term is a difficult task, and one that can seriously impact the efficiency of any related retrieval functions. Disambiguation problems can be reduced by using phrase level translations (Ballesteros and Croft 1998), unfortunately it is difficult to develop high coverage phrase translation dictionaries. Parallel texts offer a valuable translation resource, unfortunately in most cases there is no parallel content available from which to establish a parallel translation resource for search queries. The final option of using MT can work effectively. However, MT systems are generally developed to work with well structured text. User queries are typically unstructured strings of search words and phrases, which can create problems for MT based translation. A further problem is that MT systems only exist for a limited number of language pairs, and it is extremely expensive to develop an MT system for a new language pair. A significant problem which can arise for all of these translation methods is the coverage of the translation dictionaries. If the user enters queries which use words or phrasal expressions outside the translation dictionaries, they will not be translated accurately. Errors in translation arising due to ambiguity, linguistic structure or dictionary coverage are the main source of degradation in retrieval effectiveness for CLIR. Methods are currently being investigated which allow general domain translation resources to be augmented with domain specific bilingual dictionaries automatically extracted from resources such as Wikipedia (Jones et al. 2008).

4. Merging CLIR and AH

While AH and IR share many objectives in satisfying user information needs, to date they have developed almost entirely in isolation. Consequently they largely use different technologies, and have different strengths and weaknesses. One of the key research themes within the DCM track of the CNGL is the hybridization of AH and IR technologies to address user information needs more effectively. For example, statistical methods used in IR typically return a set of ranked documents, however for some applications this may not be the most user-friendly method of accessing and presenting data, nor is the selection of a complete document necessarily a suitable answer for some information needs. On the other hand, the substantial need for metadata associated with AH techniques, as well as the need for content to be structured in a particular way, means that content needs to be authored manually for a specific system. Thus we are interested in exploring the introduction of personalisation features into IR, and to introduce IR techniques to AH, with the objective of satisfying a wider variety of users and their different information need types, over a variety of languages without requiring the addition of large amounts of manual metadata.

The first mechanism that we are investigating is the combination of personalisation and IR to improve multilingual query expansion. Using structures such as a domain model and user model, the system can make determinations about the subject domain of interest for a particular user and their queries. The presentation and selection of particular results from a personalised, expanded query can also be altered using AH techniques. For example, personalisation can be used to re-rank results, for example for previously unseen results, or for results in a particular format. Currently, the activity in this area being undertaken by the DCM track of CNGL relates to eliciting user models statistically, in order to drive improved personalised, intelligent response generation.

The second approach, which is detailed below, uses IR techniques to 'crawl' the WWW and selected digital content repositories to generate collections of content in particular subject domains. This open corpus content model can be used to make large quantities and varieties of web-sourced content more accessible for incorporation in adaptively composed presentations. Several substantial challenges remain, and the following sections outline some of the research which has been undertaken in the area of transforming unstructured data into AH content.

4.1 Open Corpus Content

Enormous volumes of content, which is varied in structure, language, presentation style etc., and is suitable for inclusion in AH presentations can be sourced via the WWW. This "open corpus" content, can be defined as any content that is freely available for non-commercial use by the general public or educational institutions. Such content can be sourced from web pages, scholarly research papers, digital content repositories, forums, blogs, etc.. While some AH systems, such as KBS Hyperbook (Henze 2000) and SIGUE (Carmona 2002), allow the manual incorporation of individual web-based content resources, the scale of open corpus content available is yet to be comprehensively exploited by AH.

In order to facilitate the large-scale utilisation of open corpus content in AH, methods of surmounting the heterogeneity of web-based content must be developed. This includes integrated means of content discovery, classification, harvesting, indexing and incorporation. The Open Corpus Content Service (OCCS) (Lawless et al. 2008) is an IR tool chain which has been developed to address these challenges.

The process of content harvesting for AH is infeasible at runtime, since it requires a considerable amount of resources in order to discover and index the large volumes of data available via the WWW for a particular domain. A persistent document cache is therefore created, in order to ensure reliable content candidate selection during the creation of an AH composition. This cache ensures that: the content is permanently available, the content accurately reflects the index representation of the resource and that further content preparation can be conducted on the resources before incorporation into an AH composition.

Focused web crawling enables the discovery of content which meets pre-determined classifications from across the WWW. The OCCS applies focused crawling techniques to traverse the WWW and centrally collate open corpus content resources, categorised by subject domain, for use in AH compositions. The OCCS combines an open source web crawler called Heritrix [Heritrix] with an open source text classification library called Rainbow [Rainbow] to conduct focused crawls where discovered content is compared to a statistical model of a subject area to estimate relevancy.

Once content has been discovered and harvested it must also be indexed to make it more readily discoverable during the content candidacy process. There are numerous open-source content indexing solutions available, such as Lemur [Lemur] and Lucene [Lucene]. Some indexing tools such as Nutch [Nutch] and Swish-e [Swish-e] have also been integrated with a web crawler to form openly available information retrieval tool chains. However, these tool chains typically utilise general purpose, rather than focused, crawling techniques and can be limited by the indexing methods employed. The OCCS combines its focused crawling functionality with an open source indexing tool called NutchWAX [NutchWAX] to implement web-scale subject-specific content discovery and indexing. NutchWAX enables the indexing and free text search of web archives, or collections of web-based content.

4.2 On-Demand Slice provision from Subject-specific Caches

The OCCS delivers an integrated means of content discovery, classification, harvesting and indexing which can be leveraged by AH systems. However, there remain several challenges associated with the incorporation of crawled content in an AH composition. The first and most obvious is that conventional assumptions regarding the granularity, format and presentation style of the content available to an AH system can no longer be made. AH systems have traditionally operated upon closed sets of content resources, where the system is aware of each individual resource, its format and characteristics and any relationships between resources in advance. When attempting to incorporate open corpus content, the AH system can no longer assume it possesses this detailed knowledge.

There tends to be an inverse relationship between the potential reusability of a resource and its granularity. The larger and more complex a resource, the more contextually specific it is and the more difficult reuse becomes. Fine-grained, conceptually atomic, resources are much more easily reused outside of the context for which they were created.

The OCCS delivers resources which have been harvested from the WWW and still contain contextually-specific information such as navigation bars, advertisements, banner images etc.. These resources can also be large, course-grained pages of text. To address the challenge of improving the reusability of the resources delivered by the OCCS, research is underway into the development of a framework which can serve requests for specific content from AH systems with precise resources which have been stripped of ancillary information (Levacher et al. 2009). This would allow the AH system to request resources which conform to specific granularity, format and presentation style requirements. This would solve some of the problems with seamlessly incorporating open corpus content into AH compositions.

This framework, which implements content analysis and on-demand resource generation, is outlined in Figure 2. Each component of the framework executes a specific task on the open corpus content. These tasks include: the structural disaggregation of the content to remove presentation-specific content and other extraneous content; a statistical analysis of the content to map the concepts detailed in different parts of the resource; and the on-demand resource

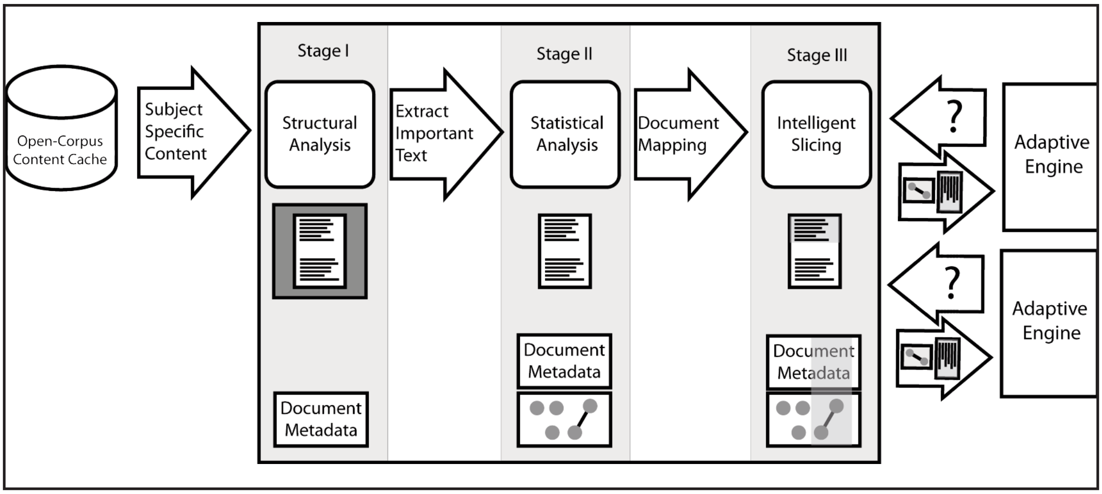


Figure 2: On-Demand Slice Generation Framework. The content to be analysed is drawn from an OCCS cache and is structurally and statistically analysed to extract the informative content and a conceptual map of that content.

provider, which fulfils requests from the AH system for specific pieces of content.

The resource delivered to the AH system by this framework is referred to as a 'slice'. A slice is an abstract notion representing a conceptually atomic piece of information, originally part of an existing resource, which has been extracted to fulfil a specific information requirement. A slice can potentially be a composition of other slices from a number of resources. Appropriate metadata is also generated for the slice based upon the metadata of the parent resource(s). A slice is virtual in the sense that it only represents a subjective perspective of a particular resource and its description. The degree of complexity of a slice will match the requirements of the AH system which has requested it.

The content analysis stages of the framework are executed a-priori while the intelligent slicing task is executed at run-time. Each a-priori framework component generates a specific layer of metadata based upon the tasks and content analysis it has conducted. This metadata augmentation enriches the resource with any structural and semantic information which has been identified during the content analysis.

The first content analysis stage is structural segmentation, which is used to remove presentation-specific content and other extraneous content. Structural segmentation techniques include densitometric analysis (Kohlschütter 2009), DOM tree pattern analysis (Vieira 2006), isotonic regression (Chakrabarti 2007), vision-based techniques (Baluja 2006) (Cai 2006) and token-based approaches (Pasternack 2009). Once the structural segmentation has been performed, a statistical analysis of the resulting content is performed to create a conceptual map of the content. Statistical analysis approaches include the use of supervised learning techniques such as hidden markov models, dictionary and rule-based approaches and word-sense clustering. These stages allow the framework to reduce the resource down to its informative content, and identify what concepts are dealt with by the resource, and at what points in the content.

This framework represents a novel approach to content candidate provision in AH systems, as it does not attempt to identify the required conceptual coverage, granularity or format of a resource in advance of presentation composition. Instead, the framework constructs a custom resource to fulfil the specific requirements of a request provided at runtime by the AH systems. The Framework architecture is currently under implementation within the DCM track of CNGL, with particular focus on the selection of structural analysis strategies.

5. Applications to Localisation

The localisation process can be viewed as a complex workflow of participants with different roles. Beginning with the content creator, the content itself must be translated into a target language and altered to conform to suitable cultural norms for a particular locale. The DCM track of CNGL aims to develop techniques which can be used to support a wide variety of localisation roles, beginning with the original author of the content, progressing through the translator who transposes the content to a new language, and ultimately to the reader who consumes the content.

There is considerable room for innovation in this area, both in the direct use of AH and IR techniques to improve content retrieval and presentation for the user, and in the use of these technologies to create greatly improved tool support for the participants in the process of generating localised content. This innovation can be viewed as affecting each of the different key roles in the localisation workflow, from the content creator to the localisation manager, Each role shares some concerns and has individual problems which can be addressed with improved digital content management.

5.1 In Support of the Content Creator

Translation reuse appears to be a common component of the overall goal of improving localisation effectiveness and efficiency. One approach to this goal is to make use of fuzzy string matching to retrieve previously-executed translations from a Translation Memory. In interviews with localisation service providers, one of the key challenges to this reuse is the tendency for content creators to seek to generate original content, which varies from writing guidelines.

There are two ways which DCM research can support improved content authoring. The first is to help in the generation of digital content, and the second is in creating tools which help authors to better comply with writing guidelines by making them more accessible.

Open-corpus AH techniques could be used to create subject-specific caches which contain information either from the open web, or from designated content sources. This can be applied to the creation of digital media through the reuse of 'slices', which can be re-composed in the formation of new documents, as described in section 4.2. Instead of a content author having to generate a completely new document, or substantially edit an existing document, it is potentially possible for an on-demand slicing service to create customised pieces of existing translatable content for use within a new document. There are several advantages to this approach over the manual authoring and merging of content, as the best content from the documents chosen over the whole corpus can be used, with the author able to concentrate on their combination rather than the laborious process of searching and separating slices manually.

The content that is available to create the appropriate slices can be selected and prepared in advance, ensuring that the new document contains only appropriate material. Conventional document merging is a difficult and labour-intensive process, because there is a need to navigate and retrieve the correct content, then read large amounts of content to select appropriate parts for combination. This can make determining the provenance of pieces and the consistency of the document text more irregular. On-demand slicing can avert this difficulty by recording the transformations applied to a particular piece of a document during the content preparation workflow as well as its origin. The use of adaptive techniques to help select and compose the slices means that an AH system backed by an on-demand slicing service can provide content authors with a coherent skeleton of a document, rather than a collection of unrelated fragments.

By improving reuse at the document generation level, it is intended that the new documents, created during this process, will be more amenable to translation reuse, and will better comply with content creation guidelines because the constituent content is chosen from the corpus of documentation which best suits the localisation process.

The second method for supporting content creators is to make it easier for them to access and understand the policies which govern their authoring. As discussed above, AH systems can be used to select content which is most relevant to a particular user model. These techniques can be used to help content authors by tracking the documents that they are working on, and presenting only the relevant portions of particular policy documents. Familiar policy elements can be summarised, while new or particularly important translation and localisation guidelines and policies can be highlighted on a personalised basis.

This allows the author to have constant access to a tailored presentation of the translation and localisation guidelines and policies that they need to be aware of for a particular document, which is presented in a style that suits their preferences, and remembers what they are familiar with. This is intended to help authors both by reminding them automatically of relevant policies, and also by reducing the overhead in switching between different documents, which might be governed by different policies. AH has been shown to be effective in classroom learning situation (Conlan 2004) and the intention is to evaluate its effectiveness in this domain.

The effectiveness of these approaches will depend on several factors. The first, and most important, is that there is a need for the tools, which are intended to support content authors, to be effective in the generation of content which communicates effectively and can be translated at lower cost. There is a need to ensure that content preparation frameworks do in fact make the creation of content more effective, and that the adaptive documentation tools are in fact useful and deliver relevant instructions. The only way to measure this is with user trials, and the intention of our work in CNGL is to evaluate DCM research outputs in industrially-relevant scenarios with industry professionals. A combination of evaluation metrics will be employed that include industrial and academic concerns.

5.2 In support of the Translator

A human translator has several roles in the treatment of content. The first, and most obvious, is the creation of new translations for content. Depending on the translator's familiarity with the content, and their domain expertise, there can be substantial skill required in correctly translating the meaning of content in the correct sense.

The first application of DCM research in this area is the use of CLIR to help retrieve background material for translators from the open web. CLIR is effective in that it will return results of documents in a variety of languages, which can help the translator by providing them with similar texts in the original and target languages. This can help the translator to improve the quality of their translations by giving them access to background material on a particular topic in all the languages that they need.

The second major task of the translator which can be supported by DCM is in the creation of adaptive courses which help with guidelines, in the same way as to support content authors. These adaptive presentations are not limited to the rules and guides governing the translation and localisation process, but can also present relevant portions of terminological dictionaries, and can even make some simple semantic inferences from the domain models to help choose terminological hints specific to the translator, and the content which they are working on.

5.3 In support of the Reader

One of the key objectives of DCM research is to improve the presentation of content to the end user. In particular, adaptive presentation of information improves the ability of users to explore content, and encourages them to examine background material and related topics (Steichen 2009). Adaptive presentations based on a personal information need are of most benefit when there is a wide selection of candidate content appropriate to supporting different personalised requirements.

Traditionally, where 'personalisation' has simply meant offering users manual choices of language and locale, the content to be presented to the user has been prepared on a mass basis into specific culture/language combinations. As personalisation evolves to adaptivity, a more refined notion of how content can vary for individuals emerges.

Supporting fully personalised content in Localisation will likely require some changes to the assumptions about how content is transformed and managed during the localisation process. The composition of content pieces could allow for a more effective management of localisable material by allowing translation resources to be concentrated on the pieces of content which are most used in the compositions, while less important content can be relegated to machine translation, which can be less expensive.

A key advantage to the fact that personalisation systems can record a model of the user's behaviour and preferences is that this model can itself be reused. For example, as a user passes between different managed content sites, the user model can be allowed to propagate across these sites, cross-influencing each site which was independent up until that point. The notion of non-invasive adaptation means that personalisation in this case can have a subtle, but nonetheless powerful effect on the way content is presented, and the ease of the user's navigation (Koidl 2009).

Much of our work is at the early stages of planning and design. As the technologies mature, it is our belief that working with localisation professionals in the field will yield technological progress and mutually beneficial results through improved digital content management.

6. Conclusions

This paper has presented a brief introduction to the area of DCM research, through the specific subjects of AH and IR. Particular focus has been placed on multilingual, open-corpus techniques, as these are likely to be the best suited to the multi-cultural WWW of the future. The application of these technologies to localisation, and their evaluation within that domain is intended to create real improvements in the way content is prepared for presentation to the user.

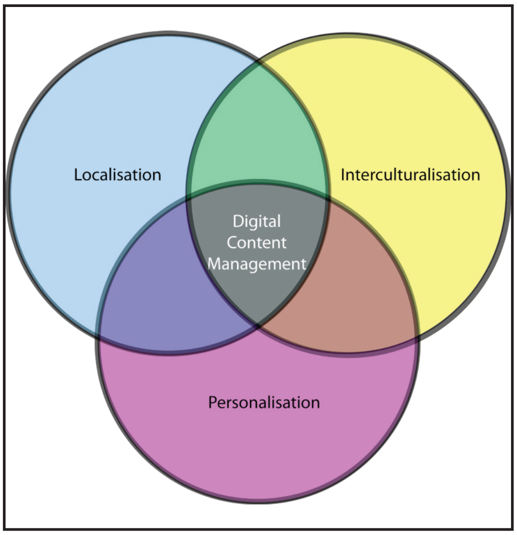


Figure 3: Locale, Culture and Personal Preferences combine to influence the future of DCM

Improvements in multilingual digital content management will result from a better understanding of the key influences which make for better responses to user's information needs. The first influence is the localisation of the document to the appropriate language. This is a fundamental requirement; it is difficult to envision a user being able to make use of a document which is not in the appropriate language. Translation of content can come in a variety of forms, and for a variety of costs, so the goal of reuse and targeting of effort seen in AH is of similar if not greater proportions in the localisation area.

The second influence of cultural appropriateness is deeply interlinked with the separation of presentation and content. There is a key need to be able to communicate content to the user with attractive and appropriate cultural norms, and once again, compatibility with the objective of DCM technologies, which seek to abstract the content of a document from surrounding extraneous features of a page.

The influence of personalisation is apparent, and has, until now, not been addressed in localisation. There are many advantages to being able to create specific content for individuals, and it allows targeting of effort across the content management workflow. The analytical aspects of DCM technologies also allow for statistics on the effectiveness and usefulness of particular content to be gathered, which can be fed back into the content preparation and analysis process for improved results.

Finally, as a rich media environment with clear performance metrics and an inherently multilingual approach, the localisation industry itself has the potential to benefit substantially from the addition of DCM techniques to improve the retrieval and presentation of the media that supports the process.

References

Adriani, M. "Using statistical term similarity for sense disambiguationin cross- language information retrieval". Inf. Retr., 2(1):71-82, 2000.

Ballesteros, L. & Croft, W.B. (1998) "Resolving ambiguity for cross-language retrieval", In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in Information Retrieval, pp 64-71, Melbourne, Australia

Baluja, S. (2006) "Browsing on Small Screens: Recasting Web Page Segmentation into an Efficient Machine Learning Framework". In the Proceedings of the 15th International World Wide Web Conference, WWW2006, pp. 33-42, Edinburgh, Scotland.

Brin, S. and Page, L. "The anatomy of a large-scale hypertextual web search engine". Comput. Netw. ISDN Syst., 30(1-7):107-117, 1998.

Brusilovsky, P. (1996) 'Methods and Techniques of Adaptive Hypermedia',User Modeling and User Adapted Interaction. 6 (2-3) : 87-129
Brusilovsky, P. (2004), 'Adaptive navigation support:

From adaptive hypermedia to the adaptive web and beyond', Psychology Journal 2, 7-23

Cai, D., Shipeng, Y., Wen, J.R. & Ma, W.Y. (2006) "Extracting Content Structure for Web Pages based on Visual Representation". In the Proceedings of the 5th Asia Pacific Web Conference, APWeb 2003, pp. 406-417, Xi?an, China. 23rd-25th April, 2003.

Carmona, C., Bueno, D., Guzmán, E., Conejo, R. (2002) "SIGUE: Making Web Courses Adaptive". In Proceedings of 2nd International Conference on Adaptive Hypermedia and Adaptive Web Based Systems, AH2002, Malaga, Spain, 29-31 May, 2002. Lecture Notes on Computer Science, Vol. 2347. Berlin: Springer Verlag, pp. 376-379. 2002.

Chakrabarti, D., Kumar, R. & Punera, K. (2007) "Page-level Template Detection via Isotonic Smoothing". In the Proceedings of the 16th International World Wide Web Conference, pp. 61-70, Banff, Alberta, Canada. May 8th-12th, 2007.

Chen, J. and Nie, J.-Y. "Parallel web text mining for cross-language IR". In Proceedings of RIAO-2000: Content-Based Multimedia Information Access, pages 188-192, CollCge de France, Paris, France, 2000.

Conlan, O. and Wade, V. (2004) Evaluation of APeLS - An Adaptive eLearning Service based on the Multi-Model Metadata-Driven Approach. In Proceedings of the 3rd ACM International Convference on Adaptive Hypermedia and Adaptive Web Systems pp 192-195, Eindhoven

De Bra, P., Brusilovsky, P. and Houben, G.-T. (1999) 'Adaptive Hypermedia, From Systems to Framework' ACM Computing Surveys, 31(4),

De Bra, P. and Calci, L. (1997) Creating Adaptive Hyperdocuments for and on the Web. In Proceedings of the AACE WebNext Conference, Toronot, pp. 149-155

Gao, J., Zhou, M., Nie, J.-Y., He, H. & Chen, W. "Resolving query translation ambiguity using a decaying co-occurrence model and syntactic dependence relations". In Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 183-190, Tampere, Finland, 2002. ACM Press.

He, D., Oard, D. W., Wang, J., Luo, J., Demner-Fushman, D., Darwish, K., Resnik, P., Khudanpur, S., Nossal, M., Subotin, M. & Leuski, A. (2003) "Making MIRACLEs: Interactive Translingual Search for Cebuano and Hindi," ACM Transactions on Asian Language Information Processing 2(3):219-244

Henze, N. and Nejdl, W. (2000) "Extendible Adaptive Hypermedia Courseware: Integrating Different Courses and Web Material". In the Proceedings of the International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems, AH2000, pp. 109-120, Berlin: Springer-Verlag, Trento, Italy. August 28th-30th, 2000.

Jones, G.J.F., Fantino, F.,Newman E. & Zhang, Y (2008) "Domain-Specific Query Translation for Multilingual Information Access Using Machine Translation Augmented With Dictionaries Mined From Wikipedia", In Proceedings of the 2nd International Workshop on Cross Lingual Information Access - Addressing the Information Need of Multilingual Societies (CLIA-2008), Hydrabad, India, pp34-41.

Kleinberg, J.M. "Authoritative sources in a hyperlinked environment". J. ACM, 46(5):604-632, 1999.

Kohlschütter, C. & Nejdl, W. (2009)"A Densitometric Approach to Web Page Segmentation". In the Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM 08, pp. 1173-1182, Napa Valley, California, USA. October 26th-30th, 2008.

Koidl K., Conlan O., Wade V. (2009) Non-Invasive Adaptation Service for Web-based Content Management Systems (DAH2009), International Workshop on Dynamic and Adaptive Hypertext: Generic Frameworks, Approaches and Techniques, Torino, Italy.

Kwok, K.L. & Dinstl., N. "NTCIR-6 monolingual Chinese and English-Chinese cross-language retrieval experiments using pircs". In the Sixth NTCIR Workshop Meeting, pages 190-197, NII, Tokyo, Japan, 2007.

Lawless, S., Hederman, L., Wade, V. (2008) "OCCS: Enabling the Dynamic Discovery, Harvesting and Delivery of Educational Content from Open Corpus Sources". In the Proceedings of the Eighth IEEE International Conference on Advanced Learning

Technologies, I-CALT 2008, Santander, Spain. 1st-5th July, 2008.

Levacher, K., Hynes, E., Lawless, S., O'Connor, A., Wade, V. (2009) A Framework for Content Preparation to Support Open Corpus Adaptive Hypermedia In Proceedings of International Workshop on Dynamic and Adaptive Hypertext: Generic Frameworks, Approaches and Techniques, Torino, Italy.

Liu, Y., Jin, R. & Chai, J. Y. "A maximum coherence model for dictionary-based cross-language information retrieval". In Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 536-543, Salvador, Brazil, 2005. ACM Press.

McCarley, J.S. "Should we translate the documents or the queries in cross-language information retrieval?" In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics, pages 208-214, College Park, Maryland, 1999. Association for Computational Linguistics.

Nie, J.-Y., Simard, M., Isabelle, P. & Durand, R. "Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the web". In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 74-81, Berkeley, California, United States, 1999. ACM Press.

Oard, D. W. and Resnik, P. (1999) "Support for Interactive Searching in Cross-Language Information Retrieval," Information Processing and Management 35(3): pp 363-379.

Pasternack, J. and Roth, D. (2009) "Extracting Article Text from the Web with Maximum Subsequence Segmentation". In the Proceedings of the 18th International World Wide Web Conference, WWW2009, pp. 971-980, Madrid, Spain. April 20th-24th, 2009.

Oard, D.W. "A comparative study of query and document translation for cross-language information retrieval". In Proceedings of the Third Conference of the Association for Machine Translation in the Americas on Machine Translation and the Information Soup, pages 472-483. Springer-Verlag, 1998.

Oard, D.W. & Hackett, P. "Document translation for cross-langauge text retrieval at the university of Maryland". In The Sixth Text Retrieval Conference (TREC-6), pages 687-696, NIST, 1997.

Salton, G. & Bucklley, C. (1988) "Term weighting approaches in automatic text retrieval" Information Prcocessing & Management, 24(5):513-523.

Spärck Jones K Walker S. & Robertson S. E. (2000) "A probabilistic model of information retrieval: development and comparative experiments Part 1". Information Processing & Management 36(6):779-808.

Spärck Jones K Walker S. & Robertson S. E. (2000) "A probabilistic model of information retrieval: development and comparative experiments Part 2". Information Processing & Management 36(6):809-840.

Steichen, B., Lawless, S., O'Connor, A., and Wade, V. (2009). Dynamic hypertext generation for reusing open corpus content. In Proceedings of the 20th ACM Conference on Hypertext and Hypermedia (Torino, Italy, June 29 - July 01, 2009). HT '09. ACM, New York, NY, 119-128

Vieira, K., da Silva, A., Pinto, N., de Moura, E., Cavalcanti, J. & Freire, J. (2006) "A Fast and Robust Method for Web Page Template Detection and Removal". In the Proceedings of the 15th ACM International Conference on Information and Knowledge Management, CIKM 06, Arlington, Virginia, USA. November 6th-11th, 2006.

Wray, R. (2009) 'Internet data heads for 500bn gigabytes', The Guardian, accessed July, 2009 http://www.guardian.co.uk/business/2009/may/18/digital-content-expansion

Wu, Y.C., Tsai, K.C. & Yang, J.C. "NCU in bilingual information retrieval experiments at NTCIR-6". In the Sixth NTCIR Workshop Meeting, pages 133-139, NII, Tokyo, Japan, 2007.

Zhou, D., Truran, M., Brailsford, T., & Ashman, H. "A Hybrid Technique for English-Chinese Cross Language Information Retrieval", in ACM Transactions on Asian Language Information Processing (TALIP) 7, 2, June 2008.

Footnotes

[Heritrix] Heritrix is the Internet Archive's open-source, extensible, web-scale, archival-quality web crawler project. Available online at: http://crawler.archive.org/

[Rainbow] Rainbow is a program that performs statistical text categorisation. Available online at: http://www.cs.cmu.edu/

[Lemur] The Lemur Toolkit is an open-source suite of tools designed to facilitate research in language modeling and information retrieval. Available online at: http://www.lemurproject.org

[Lucene] Apache Lucene is a full-featured text search engine library written entirely in Java. Available online at: http://lucene.apache.org/java/docs/index.html

[Nutch] Nutch is an open source web-search solution based upon Lucene. Available online at: http://lucene.apache.org/nutch

[Swish-e] Simple Web Indexing System for Humans - Enhanced (Swish-e), a flexible and free open source system for indexing collections of Web pages. Available online at http://www.swish-e.org

[NutchWAX] Nutch and Web Archive eXtensions is a tool for indexing and searching web archive collections. Available online at: http://archive-access.sourceforge.net/projects/nutch/

[CLEF] Cross-Language Evaluation Forum at: http://www.clef-campaign.org/

Integrated Language Technology as part of Next Generation Localisation

Julie Carson-Berndsen[1], Harold Somers[2], Carl Vogel[3], Andy Way[4]
Centre for Next Generation Localisation
[1] School of Computer Science and Informatics,
University College Dublin, Belfield, Dublin 4, Ireland
[2] School of Computer Science,
Dublin City University, Glasnevin, Dublin 9, Ireland
[3] School of Computer Science and Statistics,
Trinity College, College Green, Dublin 2, Ireland
www.cngl.ie
hsomers@computing.dcu.ie; away@computing.dcu.ie;
vogel@cs.tcd.ie; julie.berndsen@ucd.ie

Abstract

This paper describes one component of a large research project involving industry-academia collaboration between four Irish universities and nine Irish and multinational industry partners, all collaborating to develop 'Next Generation Localisation'. The project as a whole is described by van Genabith (2009); the current paper focuses on on the role in the project of state-of-the-art language technology including text and speech processing, and machine translation (MT) in its various forms. In this paper, we describe the basic and innovative research approach to integrating language technology into the overall design. We will describe research in the areas of MT, speech technology and text analytics (TA), and ways in which these three are closely integrated with each other.

Keywords: *language technology, machine translation, localisation, standards, evaluation, speech technologies, crowd sourcing, translation memories, post-editing*

1. Introduction

As mentioned by van Genabith (op. cit.), Next Generation Localisation seeks to address current problems of increased volume, access and personalisation. Regarding volume, increased automation is the only viable approach to meet the challenges posed by the spiralling amount of material to be localised worldwide. Automation is particularly relevant to the core task of localisation, namely translation. Since its inception, the localisation industry has been highly computerised, linking up and supporting teams of human translators, localisation project managers and customers with Translation Memory (TM) and terminology management systems, electronic dictionaries, translators' workbenches and localisation workflow management and quality assurance systems. However to date, core language technologies, in particular MT, have surprisingly been incorporated only sparingly into the localisation process. This is largely due to early unrealistic expectations (on the part of the users/Localisation Industry), unwarranted promises (on the part of MT researchers and developers) and the ensuing disappointment and reluctance to invest in MT technology in the localisation workflow. Right now, this situation is changing dramatically: data-driven and machine-learning (ML)-based language and MT technologies are revolutionising MT research, achieving analysis and translation quality, coverage and robustness at a cost previously unimaginable. The proper place of MT is being investigated systematically: MT must be integrated into a translation and post-editing workflow together with human translators to tackle volume and to save costs; task-dependent configurable systems covering the complete spectrum of fully automatic raw translation, human-aided machine or machine-aided human translation with their associated different levels of translation quality can each play a role in the localisation process. Novel ML-based language technologies can automatically provide metadata annotations (labels) to localisation input to automate localisation standardisation and management. And progress in MT evaluation can measure localisation output and automatically cost localisation input. All of these factors determine the approach to MT research in this project, with the aim of advancing basic research in integrated MT-centred language technologies with the following aims:

(a) to improve MT engines to achieve increased quality and hence automation of translation in the Next Generation Localisation 'Factory', in which technologies are integrated into workflows and complex information-technology-based software systems (van Genabith 2009)
(b) to develop technology which will automatically annotate localisation input with standardised localisation metadata so as to automate localisation workflows,
(c) to develop novel MT evaluation methodologies, and
(d) to evaluate the impact of automation in the localisation workflow.

Regarding **access** and **personalisation**, we see integration of MT with speech technologies as a crucial step. Small screen and non-keyboard-based devices (mobile phones, PDAs) increasingly support affordable, pervasive, on-the-move access to globally available multilingual digital content. Novel speech interfaces will be essential to compensate for the main limitations of such devices, to provide support for 'eyes-busy, hands-busy' scenarios as well as for handicapped users (e.g. the blind). Unfortunately, to date, standard localisation has made little provision for the optimal adaptation of content to such devices, even though speech recognition and synthesis can potentially extract and provide information highly relevant to personalised delivery of and access to digital content. The present project thus supports both text- and speech-based mobile access to and delivery of multilingual information as well as personalisation of information. In order to achieve this, fundamental problems of speaker and language dependence of current state-of-the-art speech recognition systems (amplified in the multilingual, mobile, instant and online Next Generation Localisation scenario) need to be solved and a proper integration of speech and translation technologies needs to be provided: while there are striking similarities between state-of-the-art ML-based approaches to speech and MT (in terms of the underlying technologies both for statistical and example-based methods), a fully integrated speech and MT system which can share and exploit information provided by both components in a mutually beneficial way is still lacking. Basic research in speech-based interfaces has the following aims:

(a) to produce tightly coupled speech and MT technology capable of mutually and maximally exploiting information provided by each component,
(b) to reduce speaker and language dependence of speech technology through the use of linguistically motivated, ML-based hierarchies,
(c) to extract information (such as gender, age, emotion) automatically from speech input relevant to personalisation and generate personalised speech output, and
(d) to evaluate the impact of speech interfaces in the context of localisation.

All of these goals are supported by research on Text Analysis (TA), for which two core tasks are defined, namely automatic annotation of localisation data with metadata, and text classification. Reliable automatic multilingual text classification is required to tune suites of novel MT and speech processing systems to text-type and genre. Automatic labelling is required to annotate multilingual input with standardised metadata to automate localisation workflows and to annotate multilingual corpora with dependency information to induce novel probabilistic transfer-based MT systems and to provide syntactic information for syntax-boosted MT systems.

In the remainder of this paper, we will provide further details of the research work already started, and planned for the future, to address these goals. The work represents a collaboration between five groups. At DCU, the research is carried out within the Language and Intelligence Research Group in the School of Computer Science and in the Centre for Translation and Textual Studies in the School of Applied Language and Intercultural Studies. Within Trinity College, research is conducted through the Phonetics Laboratory in the Centre for Language and Communication Studies, and by the Computational Linguistics Group, part of Intelligent Systems in the School of Computer Science and Statistics. At UCD, research is undertaken by the MUSTER group in the School of Computer Science and Informatics. Languages already addressed by the MT and MUSTER groups include French, German, Chinese, Arabic, Japanese, Polish and Spanish, usually paired with English, and new personnel mean that we may address in addition Turkish, Hindi, Bengali, Irish and Irish Sign Language. Of course, not all languages or langauge pairs/directions are covered to the same degree. Work on stylistic analysis at Trinity College includes English, Danish, Finnish, Norwegian and Russian.

2. Machine Translation

A huge demand for MT exists already: web service providers process millions of requests for automatic translation every day. Until recently, the service offered by Google[1] was powered by a version of the successful Systran system, which was also behind the well-known Babelfish service[2], among others. However, this older rule-based and hand-crafted technology is in the process of being replaced by a new generation of data-driven and ML-based statistical (SMT) systems, and these technologies have dominated MT research for at least the last ten years. Nevertheless, it is being recognised that purely statistics-based MT systems will not deliver the high level of quality demanded by some applications, including some localisation scenarios. Accordingly, current research seeks to integrate better linguistic processing into statistical approaches. This is not generally felt to indicate a return to rule-based approaches, but to hybrid designs where the appropriate linguistic knowledge is extracted from corpora by statistical and other means, and integrated into MT systems which are still nevertheless driven by the basic approach that was first introduced in the 1990s, and is just now reaching a maturity based on the huge amounts of research effort dedicated to it. Our belief is that MT translation quality will be further improved through fundamental advances resulting from combining the Example-based (EBMT: Nagao 1984, Somers 1999, Carl & Way 2003) and Phrase-based SMT (PB-SMT: Marcu & Wong 2002, Koehn et al. 2003) paradigms, from the introduction of syntactic information in EBMT (e.g. Hearne & Way 2006) and SMT (Chiang 2005) to better capture global reordering, and from fine-tuning ML-based systems to text-type and genre (e.g. Ueffing et al. 2007).

A large percentage of the research in this area is dedicated to the development of core MT engines. Building on previous work by the research team, we focus on six key challenges for MT research.

2.1.Syntax-based SMT

In contrast to work which has shown that SMT performance degrades when seeded with more syntactic units (cf. Koehn et al. 2003), we have shown in previous work that incorporating models of syntax into SMT systems can improve translation quality. There are two aspects to this work:

(a) incorporating syntax in the source language;
(b) incorporating syntax into the target-language model and the translation model.

With respect to the first of these, in Stroppa et al. (2007), we demonstrated that the performance of the state-of-the-art PB-SMT system Moses (Koehn et al. 2007) can be improved significantly when context-informed features of the source language (here, neighbouring words and their part-of-speech categories) are used. These context-informed features are integrated directly into the original log-linear framework (Och & Ney 2002), while benefiting from the existing training and optimisation procedures in standard PB-SMT.

Building on this previous work, in Haque at al. (2009a) we showed that using 'supertags' (Bangalore & Joshi 1999, Clark & Curran 2004; see also next paragraph) can provide still further gains, while in Haque et al. (2009b), we demonstrated that source-language dependency information can also improve target-language output. This work is ongoing with international collaborators from the University of Tilburg, using their suite of memory-based classifiers (Daelemans & van den Bosch 2005).

With respect to the second research thrust in this track, in Hassan et al. (2006) we incorporated into Och & Ney's (2002) log-linear PB-SMT framework a novel language model based on supertags as well as a translation model whose target side included supertags to improve the BLEU score (Papineni et al. 2002) on a range of tasks, and for different language pairs. The intention behind this research thrust is to build on our previous work to incorporate target-language and bilingual constraints into the range of MT systems being developed in CNGL at DCU, in conjunction with our industrial partner IBM.

2.2. Hybrid MT systems

A second challenge is to merge SMT and EBMT into improved hybrid systems: recent ground-breaking work by Way & Gough (2005a) and Groves & Way (2005) extended previous approaches to hybrid MT by showing that combining EBMT and SMT subsentential alignments in novel 'Example-based SMT' and 'Statistical EBMT' systems improved translation quality over baseline EBMT and SMT systems. We are now porting these insights to new domains and language pairs, and integrating a novel EBMT decoder (Groves, 2007). Further novel research involves combining sets of automatically

[1] http://translate.google.com
[2] http://babelfish.yahoo.com

induced generalised templates clustered around content and closed-class words in EBMT (Brown 1999, Way & Gough 2003), where such templates are commonly used to increase coverage and translation quality, and in SMT, where such templates are not used at all. This work is being undertaken in close collaboration with our industrial partner Traslán.

A further research thrust in this area extends our previous work (Way & Gough 2005b) to build large-scale Controlled Translation systems (O'Brien 2003, O'Brien & Roturier 2007). This work is being undertaken in close collaboration with our industrial partner Symantec.

2.3. Scaling up

The third challenge is automatically scaling more linguistically sophisticated systems to larger amounts of training data. Our Data-Oriented MT (DOT) systems (Hearne & Way 2006) have already been shown, with limited amounts of appropriately annotated training data, to outperform state-of-the-art SMT systems. We are now scaling up by at least two orders of magnitude the amount of training data used by these systems, and developing novel scoring methods for DOT (Galron et al. 2009) to enable closer comparison with mainstream PB-SMT systems, especially in the parameter estimation stage (cf. Och 2003).

At the same time, we are automatically inducing the complete set of resources for large-scale probabilistic transfer-based MT from parallel texts with our treebank-based multilingual, probabilistic LFG (Lexical-Functional Grammar) parsers/generators (Cahill et al. 2004, Cahill and van Genabith 2006, Graham et al. 2009).

2.4. Tuning to text-type and genre

The fourth challenge is to tune ML-based MT to text-type and genre. Optimal lexical selection and syntactic choices can only be achieved if an MT system is tuned to a particular domain. To date, there has been surprisingly little research on tuning ML-based MT technology to particular domains, text-types or genre. Given the quality and range of training material provided by industrial partners in this research, including computer systems, security, office applications, primary and secondary legislation, and printing, we are able to tune a suite of MT systems (SMT, EBMT, hybrid and probabilistic transfer-based) to domain, text-type and genre (cf. Haque et al. 2009c), according to a classification model using classifiers from the TA research reported below, and to investigate how much supplemental genre-typical training material is required, and to investigate training on comparable, rather than full parallel text, resources. This work is being conducted with many of our industrial partners, including Symantec, Traslán, DNP, and VistaTEC.

2.5. Alignment models

Our fifth challenge is to develop novel alignment models for MT technologies based on a range of types of training data: trees, strings, dependency structures etc. This work has two main foci: word alignment and phrase alignment.

Our previous work (Ma et al. 2007, Ma & Way 2009) has shown that new models of word alignment can be developed which outperform state-of-the-art methods (Och & Ney 2003, Deng & Byrne 2005, 2006). Based on this previous work, in Lambert et al. (2009), we demonstrated that tuning word alignment on an extrinsic task (MT, here) rather than intrinsically (compared to a 'gold standard' set of word alignments) can improve translation quality on a range of language pairs and on different domains. With respect to the induction of phrase pairs, our prior work has shown that statistical models of translation can be improved by incorporating example-based source-target chunks (Groves & Way 2005, 2006), as well as pairs derived using dependency (Tinsley et al. 2008) and constituency trees (Tinsley & Way 2009). This work has recently been scaled up by two orders of magnitude in Srivastava & Way (2009), as well as extended to incorporate phrase pairs induced from head-percolation information (Magerman 1995).

Ultimately, for both word and phrasal alignments, we would like to develop a novel general alignment model which, in abstracting away from the surface differences in annotation, is capable of inducing subsentential alignments over source-target pairs no matter which type of annotation is provided.

2.6. Evaluation

The final challenge in this part of the research programme is to develop improved automatic MT evaluation technology. This is a key component of MT research and development. To date, most automatic evaluation methods are based on string matching (e.g. BLEU, Papineni et al. 2002), and have been shown to penalise legitimate lexical and syntactic variation (Callison-Burch et al. 2006). In order to account for such variation, such methods require multiple references, which are time-consuming and expensive to construct. Based on our previous work (Owczarzak et al. 2006), we are developing a novel dependency-based MT evaluation technology that automatically accounts for both lexical and syntactic variation. In He & Way (2009a), we showed that the labelled dependencies in Owczarzak et al. (2006) could be learned by ML techniques with a corresponding increase in correlation with human judgements. In He & Way (2009b) meanwhile, we demonstrated that parameters trained on one metric (BLEU, say) using the method of Och (2003) may not lead to optimal scores on the same metric, especially where only a single reference translation is provided. Furthermore, combining different evaluation metrics not only reduces any bias related to any one particular metric, but also gives better translation quality than tuning on any standalone metric.

Putting all these together, we aim to produce a suite of novel core MT engines for the purposes of localisation as widely understood by the project as a whole, including PB- SMT systems with integrated syntactic models, improved models of EBMT, novel hybrid data-driven EBMT-PB-SMT systems, novel discriminative and controlled EBMT engines, large-scale DOT systems, novel transfer-based MT engines induced from automatically labelled parallel corpora, systems tuned to text-type and genre, novel aligners, EBMT decoders, and automatic MT evaluation methods.

3. Speech technologies

Flexible, non-keyboard-dependent, on-the-move voice access and response is a core enabling technology for intelligent access to digital content. Speech interfaces to mobile devices are essential in 'eyes-busy, hands-busy' scenarios. In the multilingual application scenario addressed by the CNGL research project, a tight integration of speech technologies and MT is imperative to achieve optimal results. In order to address the project goals of volume, speech recognition and synthesis systems need to deal with potentially an unlimited vocabulary, with multiple (and non-native) speakers and with multiple languages. Speech carries information on multiple levels. For example, personal information such as gender and age is communicated by voice characteristics; prosody and voice quality carry crucial grammatical information; emotional state or mood is communicated by prosody and tone of voice; sound qualities and systematic patterns distinguish between native and non-native users. Such factors are required to support personalisation.

Within the context of the Next Generation Localisation project, the key research challenges for speech technology are the design, development and evaluation of a new breed of robust and scalable automatic speech recognition (ASR) and speech synthesis engines which overcome problems associated with unrestricted domains and speaker types and which facilitate porting of the technologies to new languages (Carson-Berndsen & Walsh 2005). We are developing intelligent engines which utilise the multiple levels of human expressive speech, which are being integrated in a novel way with the MT models described above to facilitate speech-to-speech MT, text-to-speech MT and speech-to-text MT. By defining the experimentation domain for tight coupling in terms of an annotation hierarchy of linguistic information, processing in both the MT and the speech technology domains can avail of structured information at various points in the hierarchy and thus utilise both top-down and bottom-up information (cf. Figure 1).

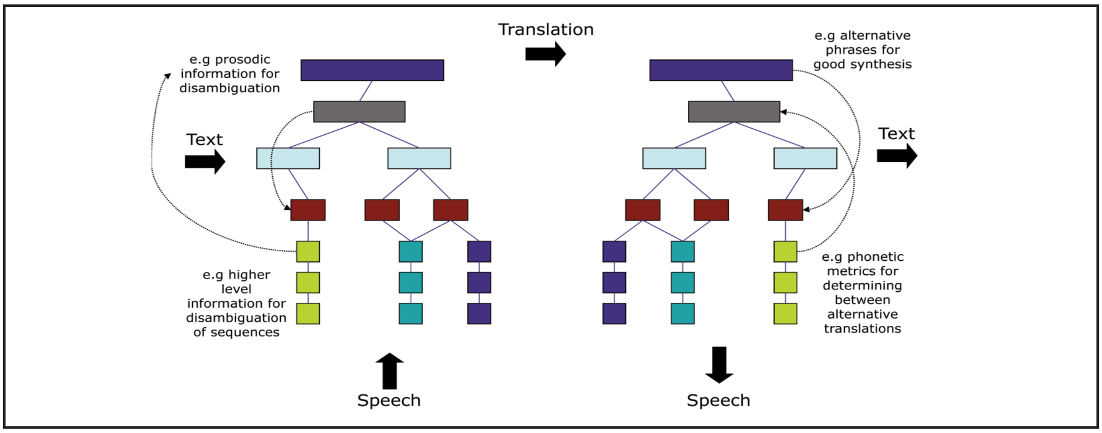

Translation
e.g prosodic information for disambiguation
e.g alternative phrases for good synthesis
Text
Text
e.g higher level information for disambiguation of sequences
e.g phonetic metrics for determining between alternative translations
Speech
Speech

Figure 1. Potential benefits of integrated language technologies

Text-to-text MT may also use the speech resources where this information can support disambiguation. We envisage the MT and speech technology constraints working together to come up with the best solution, where gains outweigh any disadvantages in terms of added complexity. Building on our previous work in speech synthesis (Cahill & Carson-Berndsen 2006), and motivated by a concatenative approach to synthesis, the synthesiser is designed to learn automatically a new language from speech data and a pronunciation dictionary. This novel approach includes an adaptive, motivated cost function which uses phonetic insights and phonotactic and morphotactic information on pronunciation and speaker variation facilitating explicit integration with the MT engines described above for use in assistive automatic interpreting. Phonetic insights on voice characteristics are being used to develop a synthesis engine that models specific emotive state to endow synthetic speech with a more human, expressive character (Gobl & Ní Chasaide 2003).

For the ASR engine, we are building on the MuSE speech recognition system, developed by Kelly et al. (2007). This system combines flexible, robust feature-extraction engines with a syllable-recognition component based on language-specific finite-state phonotactic models of speech to overcome the problems of speaker and language dependence (Carson-Berndsen 2000, Aioanei et al. 2005). We are extending MuSE using ML techniques to develop feature-extraction engines to detect phonetic characteristics of speech that are relevant across many languages (Kanokphara et al. 2006). Phonetic feature-extraction engines produce multilinear representations of speech utterances and provide a way of modelling and investigating variability; feature-based inheritance hierarchies provide information on well-formed segments and deal with underspecification. Each of the ASR engines (feature, phoneme, syllable, word) will seek to integrate explicitly with, and utilise, higher-level linguistic information from MT and TA (cf. Figure 1).

The specific outcomes of this research will be an intelligent speech-synthesis engine integrated with the MT engines described above based on a novel unit-selection approach using hierarchies of linguistically motivated units together with phonetic insights to generate natural-sounding speech. We will also develop an intelligent ASR system, integrated with MT engines, consisting of a set of feature-extraction engines and a linguistic recognition model which takes the parallel feature streams as input and outputs orthographic word sequences and annotations required by MT and TA. The speech synthesis and speech recognition engines will enable speech interfaces and facilitate access and personalisation for eyes-busy, hands-busy localisation applications. The engines will be evaluated not only with respect to standard speech synthesis and recognition metrics but also in the context of demonstrator applications with MT and adaptive content.

4. Text Analytics

The key research challenges addressed in this area are the design, development and evaluation of multilingual text-type and genre classifiers for the purposes of localisation, the automation of localisation metadata annotation to support localisation workflows, and the automatic dependency annotation for syntax-enhanced SMT and EBMT engines, and novel probabilistic transfer-based MT systems.

Building on our previous work (O'Brien & Vogel 2003, Kelleher & Luz 2005, Davy & Luz 2007), we are designing, developing and evaluating a suite of multilingual ML-based text classifiers and exploring them in the context of generating localisation workflow metadata to support the automatic choice of MT engines fine-tuned to text-type and genre. Li & Vogel (2010) continue to refine classification methods. A further application of this work relates to improved semantics of user queries related to the digital content management research described by O'Connor et al. (2009). This task encompasses three distinct problems: automatic clustering, assessment of corpus homogeneity and assignment of category labels. Although these three problems have been widely studied in general contexts, applications to MT and localisation pose novel challenges in terms of scarcity of annotated data and automatic data gathering from Web sources. Active learning techniques (Davy & Luz 2007) can help tackle the first issue; measures sensitive to hyperlink structure (Kelleher & Luz 2005) can help select effective training data from online sources. In addition, building on our previous work (Cahill et al. 2004), we are scaling our multilingual dependency annotation technology to GigaWord and Web corpora to support the automatic acquisition of probabilistic transfer-based MT and syntax-enhanced EBMT and SMT engines.

Classification may focus on the level of the sentence and its constituents or at larger intersentential levels such as the level of a document or corpus. Research continues in both directions and with due attention to multilingual considerations. At intrasentential levels, yet driven by multilingual corpus based needs, we have addressed syntactic analysis in LFG. Considerable activity at DCU has built upon a substantial platform of probabilistic parsing and generation technologies developed there and in concert with the LFG and MT communities (Bryl et al. 2009, Graham et al. 2009). A great wealth of linguistic resources in the form of treebanks have already emerged. On the semantic annotation side we have been addressing word-sense discrimination and labelling of the semantic roles that arguments fill with respect to predicates (Li et al. 2009). Language guessing is an established application of text classification techniques (Cavnar & Trenkle 1994) and in scenarios made possible by success within NGL research, multilingual chat for example, the role of language guessing for individual sentence-level contributions for the purposes of directing them to particular MT engines is possibly more clear than for high-volume document localization. However, in the industry context, automatic classification of document components on the basis of style and content, as relevant to the legal section versus hardware requirements section versus operational use sections is also relevant in the process of speeding texts on to appropriate translators. Our research into stylistic classification has use with respect to assessment of source-language conformity with house styles and general stylistic homogeneity, evaluation of translation outputs, identification of effects of source language and translator on translations, and the effects of interaction on language production. A range of scientific questions and practical applications are explored.This research has several aims:

(a) to develop and evaluate multilingual text classifiers for Next Generation Localisation;
(b) to develop and evaluate ML-based localisation workflow metadata annotators; and
(c) to develop and scale automatic multilingual dependency annotation technology to support automatic acquisition of novel transfer-based probabilistic MT and syntax-enhanced SMT and EBMT engines.

We seek to quantify theoretical bounds on the effectiveness of methods on the basis of internal properties of data under scrutiny. The research aims to exploit text classification methods in a range of scenarios that present purely theoretical problems and theoretical problems that have practical industrial relevance, not only those alluded to above.

5. Crowd-sourcing and integration

In close engagement with our industry partners, we have identified three significant emerging developments in the localisation landscape. These are:

(a) crowd-sourcing and community platforms;
(b) integration of MT systems and TMs; and
(c) interaction of MT and translation post-editing (both manual and automatic post-editing).

5.1. Crowd-sourcing

Over the last two years crowd-sourcing and the supporting community platforms have begun to take centre stage in localisation. Google, Facebook, IBM, Microsoft, Adobe, Symantec and Sun (amongst others) have successfully involved users (and customers) in a variety of localisation-related efforts, ranging from allowing users to correct the output of SMT systems and using the corrections to re-train the MT systems (Berlin 2009, Cohen 2009), to fully involving users in the localisation of web-pages and interfaces (Hosaka 2008). Crowd-sourcing is attractive as it can reduce the cost base associated with localisation activities (users do it for 'free'), and help tune the localised output to the (linguistic) requirements and expectations of a particular user (or 'fan') base. Facebook for example, having involved users to localise their social networking sites into French, German and Spanish, has reported an increase from 52M visitors to 124M as a result (Britton and McGonegal 2007:80; Eskelsen et al. 2008:120), while IBM report[3] that in the first year of its launch, 3,000 employees contributed 36M words' worth of translations. On the other hand, translators have complained that crowd-sourcing devalues their profession, when they are expected to work for free or for payment in kind (Newman 2009).

Important issues in crowd-sourcing are quality control and text domain. Facebook, for example, uses translator rankings on their sites to publicise and hence reward user-supplied translations, voting-based translation (to eliminate 'bad' translations) as well as professional translators to validate and post-edit user-localised sites before going live. Berlin (2009) mentions the need for "review by a second

[3] http://www.research.ibm.com/social/projects_nfluent.html

translator before publication and [to] have translators sign their work, discouraging sloppy or deliberately malicious translations". It is said that users are only interested in translating customer-facing content (i.e. the main web-pages visible to a Facebook client) and (unsurprisingly) are not interested in translating technical or legal documentation pertaining to the sites, which are localised using fully professional localisation operations.

Crowd-sourcing can be used to provide essential localisation services to NGOs and Development Agencies where traditional ROI calculations do not apply. Our current research involves Translators Without Borders[4] (as a user base), an organisation which provides free localisation services to NGOs such as Médecins Sans Frontières and Ashoka. Crowd-sourcing is fully integrated with the technology developments relating to MT and speech technology (see Figure 2): with access to the MT systems for post-editing, the resulting corrections will be used to retrain the MT systems.

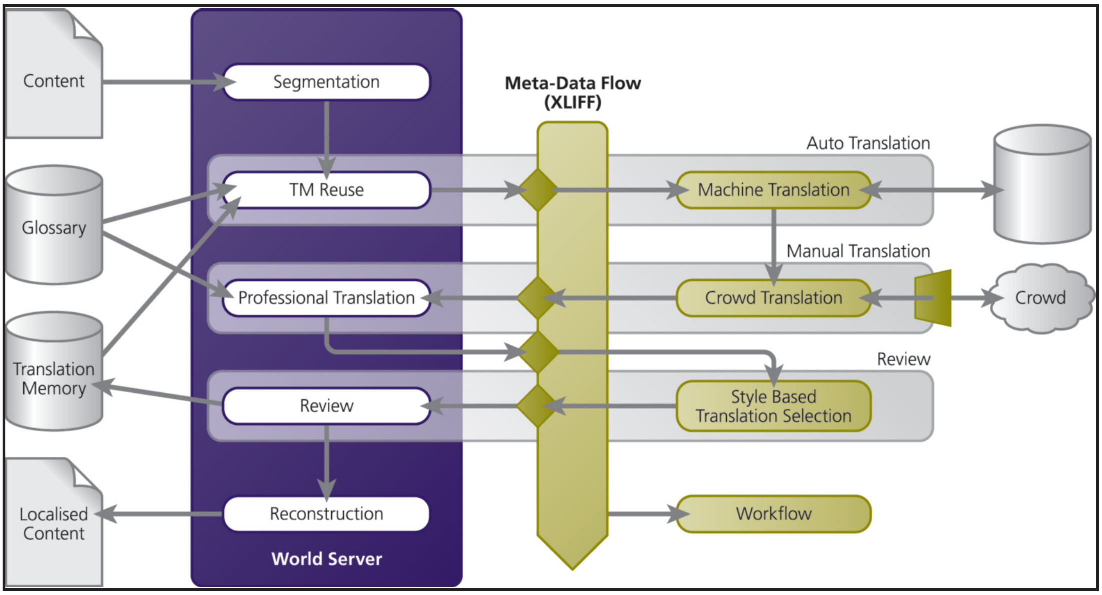


Figure 2. Localisation workflow integrating crowd-sourcing

5.2. Translation Memories

TMs are a core technology in state-of-the-art localisation workflows (Schäler 1996). In a sense, a basic TM constitutes a (very simple) EBMT system which operates at the level of complete sentences (Somers & Fernández Díaz 2004). Usually TM systems are enhanced through partial match facilities, if a complete match cannot be found. In general, partial matches come in two forms: (i) partial matches on the sentence level and (ii) complete matches on subparts of the original sentence. The first supports retrieval of translations for sentences that are similar to (but not the same as) the original sentences and the second allows matches on parts of the input string and retrieves potentially useful translations for those fragments. In each case, the TM flags the translations retrieved from partial matches to the human translator, who will post-edit the output proposed by the TM. Leveraging TMs can substantially reduce translation costs and is particularly effective for predictable text types (including technical documentation, user manuals, help files etc.).

With increased availability and continuously improving output quality, MT is beginning to make strong inroads into localisation operations. In principle, TM and MT technologies are complementary: TMs provide maximal quality performance on seen material, while MT is likely to perform better on unseen data. Combining the complementary strengths of TMs and MT approaches in the localisation workflow promises efficiency and

[4] http://tsf.eurotexte.fr/index-en.shtml

quality gains over and above the exclusive use of either technology in isolation. The challenge is provided by data in between, i.e. data (sentences) consisting of a mix of seen and unseen components. Depending on the text type, this type of data can be the most frequent. To date, the optimal combination between TM and MT technology is not known (cf. Simard and Isabelle 2009, for one recent view on this). In our research we will parameterise the problem according to a number of important (controlled) variables, including:

(a) MT type: rule-based vs. data-driven;
(b) text type: predictable vs. unpredictable;
(c) MT output quality: good vs. poor;
(d) language pairs: closely or distantly related and
(e) well- or poorly-resourced.

We expect that the space defined by these variables will lead to different optimal combinations of TM and MT technologies in the workflows. For example, MT in its example-based form can be seen as an extension of a TM in that while using a TM it is up to the human translator to decide what to do with the proposed match, in EBMT the system takes the match and tries to manipulate it to provide a translation. One way in which the performance of a TM can be improved is where a match differs minimally from the text to be translated: by connecting the TM to target-language resources such as parallel corpora or even simple dictionaries, the target-language text to be changed can be highlighted, and a translation proposed - a rudimentary type of EBMT. Another example is that with data-driven MT systems (such as SMT) we can automatically generate TMs from the training sections of the SMT system and use the resulting TM/MT combination in the space defined above; in the other direction, we can use the aligned text in TMs and train an SMT system.

5.3. Post-editing

Manually post-editing MT output is an emerging task in localisation workflows. However, to date, translators are rarely trained, if at all, to post-edit MT output - a quite different task from revising human translations (Loffler-Laurian 1985, McElhaney & Vasconcellos 1988:141, Allen 2003) - and computerised translation tools (such as editors in translators' workbenches) do not specifically support post-editing tasks. In fact, post-editing is often viewed highly unfavourably by professional translators. Our research investigates post-editing strategies in localisation scenarios, to develop recommendations for how a post-editing interface would support the observed strategies and identify training needs for translators using MT support. We are interested in measuring correlations, if they exist, between MT evaluation metrics (e.g. Translation Edit Rate, Snover et al. 2006) or MT system-generated confidence scores and post-editing effort, measured in terms of the time taken to complete the task or the number of edits made in the segment. We are also interested in correlations between post-editing effort and linguistic features such as length of the source segment, number of verbs in the source segment, number of nouns or noun phrases etc. Additionally, our research includes investigations into correlations between years of professional experience and post-editing effectiveness.

Automatic statistical post-editing is a new and exciting area in MT. In statistical post-editing, the output of one MT system is used to train a second-stage MT system, which operates on the output of the first MT system, with the intention to improve overall translation quality. To date, such system cascades have used a first-stage rule-based MT system followed by an SMT system (e.g. Simard et al. 2007). Improvements in automatic translation quality will, of course, correspond to cost savings in localisation workflows (due to reductions in post-editing efforts in off-line scenarios and the ability to post unedited quality output in on-line scenarios).
The central research idea is to generalise this architecture to a recursive target-side cascaded self-training architecture for MT, using a variety of ML-based MT architectures.

6. Conclusion

Basic research across ILT continues in collaboration with industry partners where mutual scientific interests also have industrial relevance. The research programme has been dynamic in responding to emerging problems and will continue to do so at the same time that foundational matters in text analytics are explored.

Acknowledgement

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) at Dublin City University, Trinity College Dublin, and University College Dublin.

References

Aioanei, D., Neugebauer, M. and Carson-Berndsen, J. (2005) 'Validation techniques for parallel feature streams: the case of phoneme identification for speech recognition', Archives of Control Sciences, 15, 279-290.

Allen, J. (2003) 'Post-editing', in Somers, H., ed., Computers and Translation: A Translator's Guide, Amsterdam: John Benjamins, 297-317.

Bangalore, S. and Joshi, A. (1999) 'Supertagging: An approach to almost parsing', Computational Linguistics, 25, 237-265.

Berlin, L. (2009) 'Translating online content for love of language: Volunteers provide nuance not available with automated systems, International Herald Tribune, 18 May.

Britton, D.B. and McGonegal, S. (2007) The Digital Economy Fact Book, Ninth Edition 2007, Washington, D.C.: The Progress & Freedom Foundation.

Brown, R.D. (1999) 'Adding linguistic knowledge to a lexical example-based translation system', in Proceedings of the 8th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI 99), Chester, England, 22-32.

Bryl, A., van Genabith, J. and Graham, Y. (2009) 'Guessing the grammatical function of a non-root f-structure in LFG', in 11th International Conference on Parsing Technologies (IWPT'09), Paris, France, 146-149.

Cahill, A., Burke, M., McCarthy, M., O'Donovan, R., van Genabith, J. and Way, A. (2004) 'Long-distance dependency resolution in automatically acquired wide-coverage PCFG-based LFG approximations', in ACL-04, 42nd Annual Meeting of the Association for Computational Linguistics, Barcelona, Spain, 319-326.

Cahill, P. and Carson-Berndsen, J. (2006) 'The Jess Blizzard Challenge 2006 entry', in Blizzard Challenge 2006 Workshop, Interspeech 2006 - ICSLP, Pittsburgh, PA., [4 pages]

Cahill, A. and van Genabith, J. (2006) 'Robust PCFG-based generation using automatically acquired LFG approximations', in COLINGoACL Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Sydney, Australia, 1033-1040.

Callison-Burch, C., Osborne, M. and Koehn, P. (2006) 'Re-evaluating the role of BLEU in machine translation research', in EACL-2006, 11th Conference of the European Chapter of the Association for Computational Linguistics, Trento, Italy, 249-256.

Carl, M. and Way, A., eds. (2003) Recent Advances in Example-Based Machine Translation, Dordrecht: Kluwer.

Carson-Berndsen, J. (2000) 'Finite state models, event logics and statistics in speech recognition', in Spärck Jones, K.I.B., Gazdar, G.J.M. and Needham, R.M., eds., Computers, Language and Speech: Integrating formal theories and statistical data, Philosophical Transactions of the Royal Society, Series A, vol. 358 (1769), 1255-1266.

Carson-Berndsen, J. and Walsh, M. (2005) 'Phonetic time maps: Defining constraints for multilinear speech processing', in Barry, W.J. and van Dommelen, W., eds., The Integration of Phonetic Knowledge in Speech Technology, Dordrecht: Springer, 45-66.

Cavnar, W.B. and Trenkle, J.M. (1994) 'N-gram-based text categorization', in Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, Nevada, 161-175.

Chiang, D. (2005) 'A hierarchical phrase-based model for statistical machine translation', in 43rd Annual Meeting of the Association for Computational Linguistics, Ann Arbor, MI, 263-270.

Clark, S. and Curran, J. (2004) 'The importance of supertagging for wide-coverage CCG parsing', in Coling-2004: 20th International Conference on Computational Linguistics, Geneva, Switzerland, 282-288.

Cohen, N. (2009) 'A translator tool with a human touch', The New York Times, 22 Nov.

Daelemans, W. and van den Bosch, A. (2005) Memory-Based Language Processing, Cambridge: Cambridge University Press.

Davy, M. and Luz, S. (2007) 'Active learning with history-based query selection for text categorisation', in Amati, G., Carpineto, C. and Romano, G., eds., Advances in Information Retrieval, 29th European Conference on IR Research, ECIR 2007, Rome, Italy, LNCS 4425, Dordrecht: Springer, 695-698.

Deng Y. and Byrne, W. (2005) 'HMM word and phrase alignment for statistical machine translation', in Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Vancouver, BC, Canada, pp. 169-176.

Deng Y. and Byrne, W. (2006) 'MTTK: An alignment toolkit for statistical machine translation', in Proceedings of the Human Language Technology Conference of the NAACL, New York City, NY, 265-268.

Eskelsen, G., Marcus, A. and Ferree, W.K. (2008) The Digital Economy Fact Book, Tenth edition, 2008, Washington, D.C.: The Progress & Freedom Foundation.

Galron, D., Penkale, S., Way, A. and Melamed, I.D. (2009) 'Accuracy-based scoring for DOT: Towards direct error minimization for data-oriented translation', in EMNLP 2009, Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Singapore, 371-380.

Gobl, C. and Ní Chasaide, A. (2003) 'The role of voice quality in communicating emotion, mood and attitude', Speech Communication, 40, 189-212.

Graham, Y., Bryl, A. and van Genabith, J. (2009) 'F-structure transfer-based statistical machine trans-lation', in Proceedings of Lexical Functional Grammar 2009, 14th International LFG Conference, Cambridge, UK [2 pages].

Groves, D. (2007) Hybrid Data-Driven Models of Machine Translation, unpublished thesis (PhD), Dublin City University.

Groves, D., Hearne, M. and Way, A. (2004) 'Robust sub-sentential alignment of phrase-structure trees', in COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics, Geneva, Switzerland, 1072-1078.

Groves, D. and Way, A. (2005) 'Hybrid data-driven models of MT', Machine Translation, 19, 301-323.

Groves, D. and Way, A. (2006) 'Hybridity in MT: Experiments on the Europarl corpus', in Proceedings of the 11th Conference of the European Association for Machine Translation, Oslo, Norway, 115-124.

Haque, R., Naskar, S., Ma, Y. and Way, A. (2009a) 'Using supertags as source language context in SMT', in Proceedings of EAMT-09, the 13th Annual Meeting of the European Association for Machine Translation, Barcelona, Spain, 234-241.

Haque, R., Naskar, S., van den Bosch, A. and Way, A. (2009b) 'Dependency relations as source context in phrase-based SMT', in Proceedings of PACLIC 23: the 23rd Pacific Asia Conference on Language, Information and Computation, Hong Kong, (forthcoming).

Haque, R., Naskar, S., van Genabith, J. and Way, A. (2009c) 'Experiments on domain adaptation for English-Hindi SMT', in Proceedings of PACLIC 23: the 23rd Pacific Asia Conference on Language, Information and Computation, Hong Kong, (forthcoming).

Hassan, H., Hearne, M., Way, A. and Sima'an, K. (2006) 'Syntactic phrase-based statistical machine translation', in IEEE/ACL 2006 Workshop on Spoken Language Translation, Palm Beach, Aruba, 238-241.

He, Y. and Way, A. (2009a) 'Learning labelled dependencies in machine translation evaluation', in Proceedings of EAMT-09, the 13th Annual Meeting of the European Association for Machine Translation, Barcelona, Spain, 44-51.

He, Y. and Way, A. (2009b) 'Improving the objective function in minimum error rate training', in Proceedings of the Twelfth Machine Translation Summit, Ottawa, Canada, 238-245.

Hearne, M. and Way, A. (2006) 'Disambiguation strategies for data-oriented translation', in 11th Annual Conference of the European Association for Machine Translation - Proceedings, Oslo, Norway, 59-68.

Hosaka, T.A. (2008) 'Facebook asks users for free translations of Web site's new international versions', AP Worldstream, 18 Apr.

Kanokphara, S., Macek, J. and Carson-Berndsen, J. (2006) 'Comparative study: HMM and SVM for automatic articulatory feature extraction', in Ali, M.

and Dapoigny, R., eds., Advances in Applied Artificial Intelligence: 19th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2006, Annecy, France, LNCS 4031, Berlin: Springer, 674-681.

Kelleher, D. and Luz, S. (2005) 'Automatic hypertext keyphrase detection', in IJCAI-05, Nineteenth International Joint Conference on Artificial Intelligence, Edinburgh, Scotland, 1608-1609.

Kelly, R., Carson-Berndsen, J., Macek, J., Aioanei, D., Kanokphara, S. and Cahill, P. (2007) MuSE: The Muster Speech Engine, MUSTER Technical Report, School of Computer Science and Informatics, University College Dublin.

Koehn, P., Och, F.J. and Marcu, D. (2003) 'Statistical phrase-based translation', in Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, Edmonton, Alberta, Canada, 127-133.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin A. and Herbst, E. (2007) 'Moses: Open source toolkit for statistical machine translation', in ACL 2007 Proceedings of the Interactive Poster and Demonstration Sessions, Prague, Czech Republic, 177-180.

Lambert, P., Ma, Y., Ozdowska, S. and Way, A. (2009) 'Tracking relevant alignment characteristics for machine translation', in Proceedings of the Twelfth Machine Translation Summit, Ottawa, Canada, 268-275.

Li, B., Emms, M., Luz S. and Vogel, C. (2009) 'Exploring multilingual semantic role labeling', in Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task, Boulder, Colorado, 73-78.

Li, B. and Vogel, C. (2010) 'Leveraging sub-class partition information in binary classification and its application', in Bramer, M., Ellis, R. and Petridis, M., eds., Research and Development in Intelligent Systems XXVI, Incorporating Applications and Innovations in Intelligent Systems XVII, London: Springer, 299-304.

Loffler-Laurian, A.-M. (1985) 'Traduction automatique et style', Babel, 31 (2), 70-76.

Ma, Y., Stroppa, N. and Way, A. (2007) 'Bootstrapping word alignment via word packing', in ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Prague, Czech Republic, 304-311.

Ma, Y. and Way, A. (2009) 'Bilingually motivated domain-adapted word segmentation for statistical machine translation', in EACL 2009, Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, Athens, Greece, 549-557.

Magerman, D.M. (1995) 'Statistical decision-tree models for parsing', in 33rd Annual Meeting of the Association for Computational Linguistics, Cambridge, Massachusetts, 276-283.

Marcu, D. and Wong, W. (2002) 'A phrase based, joint probability model for statistical machine translation', in Proceedings of the 7th Conference on Empirical Methods in Natural Language Processing (EMNLP), Philadelphia, PA, 133-139.

McElhaney, T. and Vasconcellos, M. (1988), 'The translator and the postediting experience', in Vasconcellos, M., ed., Technology as Translation Strategy, Binghamton, NY: State University of New York at Binghamton (SUNY), 140-148.

Nagao, M. (1984) 'A framework of a mechanical translation between Japanese and English by analogy principle', in Elithorn, A. and Banerji, R., eds., Artificial and Human Intelligence: Edited Review Papers at the International NATO Symposium on Artificial and Human Intelligence Sponsored by the Special Programme Panel Held in Lyon, France October, 1981, Amsterdam, North-Holland: Elsevier Science Publishers, 173-180.

Newman, A.A. (2009) 'Translators wanted for LinkedIn, especially if they don't ask for any pay: Does Web site's request exploit professionals or give them exposure?', International Herald Tribune, 30 Jun.

O'Brien, C. and Vogel, C. (2003) 'Spam filters: Bayes vs. chi-squared; letters vs. words', in Proceedings of the 1st International Symposium on Information and Communication Technologies, Dublin, Ireland, 298-303.

O'Brien, S. (2003) 'Controlling Controlled English: An analysis of several controlled language rule sets', in Controlled Language Translation, EAMT-CLAW-03, Dublin, Ireland, 105-114.

O'Brien, S. and Roturier, J. (2007) 'How portable are controlled language rules? A comparison of two empirical MT studies', in MT Summit XI, Copenhagen, Denmark, 345-352.

Och, F.J. (2003) 'Minimum error rate training in statistical machine translation', in 41st Annual Meeting of the Association for Computational Linguistics, Sapporo, Japan, 160-167.

Och, F.J. and Ney, H. (2002) 'Discriminative training and maximum entropy models for statistical machine translation', in ACL-2002: 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, 295-302.

Och, F.J. and Ney, H. (2003) 'A systematic comparison of various statistical alignment models', Computational Linguistics, 29, 19-51.

Owczarzak, K., Groves, D., van Genabith, J. and Way, A. (2006) 'Contextual bitext-derived paraphrases in automatic MT evaluation', in HLT-NAACL 06 Statistical Machine Translation, Proceedings of the Workshop, New York City, USA, 86-93.

Papineni, K., Roukos, S., Ward, T. and Zhu, W.J. (2002) 'BLEU: A method for automatic evaluation of machine translation', in ACL-2002: 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, 311-318.

Schäler, R. (1996) 'Machine translation, translation memories and the phrasal lexicon: The localisation perspective', in EAMT Workshop TKE '96, Vienna, Austria, 21-34.

Simard, M., Goutte, C. and Isabelle, P. (2007) 'Statistical phrase-based post-editing', in Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics, Rochester, NY, 508-515.

Simard, M. and Isabelle, P. (2009) 'Phrase-based machine translation in a computer-assisted translation environment', in MT Summit XII: Proceedings of the Twelfth Machine Translation Summit, Ottawa, ON, Canada, 120-127.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L. and Makhoul, J. (2006) 'A study of translation edit rate with targeted human annotation', in AMTA 2006: Proceedings of the 7th Conference of the Association for Machine Translation in the Americas, "Visions for the Future of Machine Translation", Cambridge, MA, 223-231.

Somers, H. (1999) 'Review article: Example-based machine translation', Machine Translation, 14, 113-158; repr. (revised) as 'An overview of EBMT' in Carl, M. and Way, A., eds., Recent Advances in Example-Based Machine Translation, Dordrecht (2003): Kluwer, 3-57.

Somers, H. and Fernández Díaz, G. (2004) 'Translation memory vs. example-based MT: What is the difference?', International Journal of Translation, 16 (2), 5-33.

Srivastava, A. and Way, A. (2009) 'Using percolated dependencies for phrase extraction in SMT', in Proceedings of the Twelfth Machine Translation Summit, Ottawa, Canada, 316-232.

Stroppa, N., van den Bosch, A. and Way, A. (2007) 'Exploiting source similarity for SMT using context-informed features', in TMI-2007: Proceedings of the 11th International Conference on Theoretical and Methodological Issues in Machine Translation, Skövde, [Sweden], 231-240.

Tinsley, J., Ma, Y., Ozdowska, S. and Way, A. (2008) 'MaTrEx: The DCU MT System for WMT 2008', in ACL-08: HLT, Third Workshop on Statistical Machine Translation, Columbus, Ohio, 171-174.

Tinsley, J. and Way, A. (2009) 'Parallel treebanks and their exploitability in machine translation', Machine Translation (in press).

Ueffing, N., Haffari, G. and Sarkar, A. (2007) 'Transductive learning for statistical machine translation', in ACL 2007: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Prague, Czech Republic, 25-32.

van den Bosch, A., Stroppa, N. and Way, A. (2007) 'A memory-based classification approach to marker-based EBMT', in METIS-II Workshop: New Approaches to Machine Translation, Leuven, Belgium, [10 pages] van Genabith, J. (2009) 'Next generation localisation', Localisation Focus, The International Journal of Localisation, this volume

Way, A. and Gough, N. (2003) 'wEBMT: Developing and validating an EBMT system using the World Wide Web', Computational Linguistics, 29, 421-457.

Way, A. and Gough, N. (2005a) 'Comparing example-based and statistical machine translation', Journal of Natural Language Engineering, 11, 295-309.

Way, A. and Gough, N. (2005b) 'Controlled translation in EBMT', Machine Translation, 19, 1-36.
O'Connor, A., Lawless, S., Zhou, D., Jones, G. J. F, Way, A. (2009) 'Applying Digital Content Management to Support Localisation', Localisation Focus, The International Journal of Localisation, this volume

Music Localisation: Active Music Content for Web Pages

Ian R O'Keeffe
Centre for Next Generation Localisation,
Localisation Research Centre,
University of Limerick, Ireland
www.localisation.ie
ian.okeeffe@ul.ie

Abstract

Localisation is a far-reaching discipline, covering much more than just translation. Many other aspects of web site design also require attention, one example being cultural modification. Music content is often included in web sites, but how much thought goes into the cultural suitability of such music? Is the style, the genre, of the music in keeping with the locale of the user? Different cultures can also derive differing meanings and understanding from music due to their cultural conditioning, and an acceptance that the structures of Western Art Music can be viewed as universal is a dangerous assumption to make. For example, some music deemed happy in some cultures can be perceived as melancholy in others. This paper presents a novel approach for possibly 'localising' a piece of music - via a system the author originally created for the purposes of capturing and recreating emotive content in music - thus permitting the cultural modification of musical digital content held in the MIDI (Musical Instrument Digital Interface) file format. A graphical user interface allows users to alter music until they are happy with the resulting cultural or emotive content, and these presets can then be saved and re-applied to any other musical content. This approach has the enormous benefit of allowing all listeners to participate, not just skilled musicians.

Keywords: *Localisation; Internationalisation; Emotive Musicology;*

1. Introduction

1.1Research Area

Localisation affects many aspects of online digital content. From a purely translational standpoint, it deals with the translation of the language itself, and the alteration of the text with regard to its font, its directionality, and its layout. But there is more to localisation than just translation, as different cultures react differently to online content, thanks to the cultural conditioning that occurs with every member of the world population as they grow up, and experience life around them. Their expectations, their reactions, and their understanding are shaped by their culture, what language they speak, how they interact with their peers, etc. The localisation industry attempts to cater for all of these considerations by addressing many aspects of online content that go beyond the language itself and how it is presented.

For instance, colours are an important consideration, as they can have unexpected cultural implications. Take, for example, financial trading, where the colour a currency is presented in can have implications for whether it is in credit or in debit (Multilingual 2001); also, white is commonly associated with mourning in Japan (Fact Monster 2004).

Images also need to be handled carefully, as it is dangerous to assume that a symbol on a button, for example, has a clear meaning for all users. Take, for example, the icons for creating numbered or bulleted lists in MS Word. They are designed to appear correct for a left-to-right environment, such as English, and other European languages:

```
1---      or      o---
2---              o---
3---              o---
```

Most Middle-Eastern languages read right-to-left, and if the icon is to correctly reflect the functionality of its use it should therefore look more like:

```
---1      or      ---o
---2              ---o
---3              ---o
```

Images can also cause offence in some cultures if used inappropriately, such as the use of images containing hand gestures that may seem innocuous in other cultures (Jacko 2009).

But what about the localisation of digital music content? Firstly, it is necessary to evaluate the existing research in musical cognition, particularly any research that focuses on cross-cultural differences in comprehension and understanding, because if there is no evidence of such cultural diversity then the concept of music localisation becomes superfluous.

The majority of the work in this area tends to focus on the psychological aspects of musical understanding, and the physiological results, rather than on how the music itself may be adapted to suit differing cultures or locales. The research also generally takes the form of passive studies of test subjects and their reactions to pre-prepared musical data, rather than an active approach where the test subjects are able to change the music themselves. Looking firstly at psychological approaches, Gregory and Varney conducted a study that looked at the affective response of subjects to music from different cultures (Gregory, Varney 1996), and found that listeners "brought up in the Indian cultural tradition have difficulty in appreciating the emotional connotations of western music". Walker (Walker 1996) states that "understanding the music of another culture requires assimilation of the influences affecting musical behaviour as much as of the resultant musical products", suggesting that cultural conditioning plays a part in how a listener understands music. Different cultures and ethnic backgrounds also show preferences for different types of music for musical therapy, as demonstrated in a study of the music selected by medical patients to help with post-operative pain relief (Good et al 2000). Moving on to neuro-science, a study (Morrison et al 2003) of human brain activity captured using functional magnetic resonance imaging (fMRI) showed that there were activation differences between Western (familiar to test group) and Chinese (unfamiliar) music based on training. Trained listeners showed extra activation "in the right and left midfrontal regions for Western music and Chinese music, respectively". It would therefore seem that we react differently to unfamiliar musical styles or traditions whether we want to or not!

What is reassuring in the research discussed here is that there does seem to be some evidence that music is not the universal language that it is often thought to be, and that there is a place for musical localisation within the general localisation workflow. Music is fairly pervasive in terms of online web content, in both symbolic (MIDI) and waveform (.wav, .mp3 and so on) file formats, but despite this there seems to be little research in the area of music modification for cultural reasons; this helps in forming an opportunity for proposing a method or system for capturing cultural music templates, and then using these templates to localise musical data. This would then open the possibility of allowing music to adapt automatically to the locale, culture, or even the IP address, of the user.

The localisation process would have to consider two possible paths for musical translation, however, one differing widely from the other in terms of approach.

Firstly, we have the replacement approach, where the process would feature the complete removal of musical content, so it could be replaced with something more culturally suitable. Examples could include attempting to match one form of folk music with something of a similar cultural positioning in another country. This approach would require a huge database of suitable music clips, and a complex management strategy in terms of genre categorization, and cultural suitability, and would run into massive issues of storage, musical rights for performance, and data maintenance. For these reasons, this approach will not be pursued in this paper.

Secondly, there is the adaptation approach, where the existing musical content would be shaped to better match the target culture. This could be viewed as an actual translation of the digital musical content, to make it more culturally-suitable, either emotionally or categorically. This is the approach that this paper will explore here.

Of course, facilitating this kind of flexibility would require the handling of a number of considerations:

- Online musical content would need to be tagged, so that the required response from the user was established in advance. The original author of a webpage, for example, may have a particular mood in mind when selecting a piece of music, but if this is not stored somewhere as metadata then the localisation process becomes one of interpretation, which is always open to error. This tag could represent the emotion of the music required (happy or sad?), or perhaps its category or mood (upbeat, youth culture, news, corporate and so on). This tagging would be similar to the tagging of images with their literal meaning to aid in their localisation, as in this HTML example, where a *.gif image is tagged as an "Angry face":

```
<img src="angry.gif" alt="Angry face" />
```

- Once the required emotion or category was established, then the music would have to be culturally adapted to correlate with the locale or culture of the user, while also maintaining the requirements of the tag. To explain the rationale behind this requirement, consider how cultural conditioning could induce different emotional responses in listeners to the same piece of music dependant on their locale. As an example, consider a news website that tagged a 'sad' news story with a piece of music in a minor key. This piece would be accepted as sounding sad to most western listeners, but a very similar scale structure could be regarded as representing joy in some areas of the Middle East, thus causing confusion, and possible anger, if the soundtrack was viewed as a happy or frivolous counterpoint to a tragic news event.

- The next consideration, and it is a significant one, is how to alter musical content in the manner we require. The music would need to be stored in some form of notation that would allow it to be edited and parsed, and this notation would preferably be of a fairly universal nature. Then some method of editing the music automatically would need to be put in place. Finally, alteration guidelines suitable for the target culture or context would need to be applied to the piece to correctly 'localise' it for the intended user.

- The final consideration, assuming the existence of the prerequisites mentioned above, would be the creation of the musical localisation guidelines themselves. These could be viewed as a form of musical template for the required mood, emotion or category, and their selection and application would facilitate the localisation of the content they are applied to.

What this paper will present is a possible approach for handling all of these considerations, thus providing the localisation industry with a mechanism for automatically, or by choice, altering digital musical content online to match the locale or cultural requirements of the user. In terms of positioning, this approach is best viewed as being placed at the intersection of localisation, music psychology and music technology.

More specifically, this paper is placed within a sub discipline of music psychology, that of music cognition; this area is concerned with the study of music as information, from the viewpoint of cognitive science. This discipline shares the interdisciplinary nature of other fields such as cognitive linguistics. Music technology is the result of applying computers and other forms of technology to the creation and adaptation of music, and localisation here refers to the alteration of musical content to match a locale, a language, or a culture. It should be noted that the initial focus of the research presented here was on the capture of emotional content in music for one particular culture, listeners to western art music, but the data gathered in this particular application demonstrates a lot of promise for the expansion of this technical approach into other areas of music analysis and modification, and it is these potential areas of research that will be covered here.

Expanding the initial study from a western-centric bias to cover cultural differences in different locales affects two aspects of the research; on the one hand, amending the system to cope with differing musical cultures, such as dealing with different scale and pitch intervals and harmonies, and on the other, conducting studies with the existing system, but with different cultural groups as participants. Such studies could be used to capture templates for individual cultures, with these templates then being used to localise online digital content as it is encountered by the user.

The system presented here uses the MIDI music file format for storage and manipulation of the musical content. It was selected because of its wide availability, portability, and small storage footprint, and because it can be played by almost anything, ranging from PCs to mobile phones and PDAs. It does have some limitations, and perhaps other notation methods would need to be considered in the future, but MIDI is certainly a good place to start given its wide acceptance worldwide.

1.2 Background

The starting point for the research described here was an evaluation of cognitive musicology. Cognitive musicology is an interdisciplinary field of research that evolved during the 1970s from such diverse sources as cognitive anthropology, artificial intelligence, cognitive psychology, linguistics, musicology, neuroscience, psychoacoustics, speech recognition, and semiotics (Laske 1992). It initially appeared to check all the right boxes where emotive evaluation of musical content was concerned, but its bias towards a computational approach, with the system attempting to mimic human behaviour, was not what was required for collecting the emotive data required. It did, nonetheless, provide useful insights into the manipulation of musical data for the system that was eventually developed.

What was really needed was human interaction so as to enable the extraction of actual cultural indicators from music, and for this reason the research presented here uses human test subjects controlling low-level musical manipulations, and then storing these presets as a representation of their emotive preferences. This positions the system as a research tool, a technical solution for data gathering, rather than as the primary problem solver, and plays to the strengths of both human intellect and computer technology.

When specific study in the area of meaning and emotion in music is considered, mention must be made of Leonard B. Meyer's "Emotion and Meaning in Music" (Meyer 1956). This work focuses on the meaning and emotion held within music, or more specifically, within Western music's theoretical tradition, as viewed from a psychological perspective. Meyer starts out by isolating a number of contrasting viewpoints on what constitutes musical meaning and the processes involved in communicating it; firstly citing the dichotomy between absolutists, who "insist that musical meaning lies exclusively within the context of the work itself" (p.1), and referentialists, who contend that "music also communicates meanings which in some way refer to the extramusical world of concepts, actions, emotional states, and character" (p.1). He then points out that these two meanings are by no means mutually exclusive, as they "can and do coexist in one and the same piece of music, just as they do in a poem or a painting" (p.1), conceding an opportunity for establishing some common ground between the opposing arguments. He also makes the concession that his work is primarily focused on the "closed context of the musical work itself" (p.2), but that this does not rule out the importance or existence of other kinds of musical meaning, such as those introduced through cultural reference or cultural conditioning.

The aim of the research presented here is to postulate how one could capture this cultural content via a new, interactive approach, and then use this data to alter other pieces of music. It is not a musicological study of existing music, but more an empirical, computational approach that involves the subject directly in the process of creating culturally altered music. In some ways, it is analogous to the localisation process of highlighting imagery inappropriate to a certain culture, and altering it to make it more acceptable. When you consider how music is able to affect us physically, bringing us to tears, making us feel overjoyed, or energised, it definitely presents itself as worthy of careful consideration by the designers of online digital content.

The approach presented in this paper strives to give the test subject direct control over the music itself. This approach is realised through the creation and use of an interactive computer system that involves the subject directly in the process of creating the required culturally-suitable emotional states or moods through music. This is achieved by crafting a series of low-level musical operations from scratch and creating a system that allows the user to build culturally-specific modification presets with these operations, thus synthesizing the required emotion. This has the enormous benefit of allowing all music listeners to participate, not just skilled musicians, as music affects someone whether they can play an instrument or not.

By contrast, previous approaches to analyzing emotive content in music have involved the test subject as a passive listener, reacting to previously prepared material; or as a skilled musical performer, required to improvise on the spot. The former does not give the test subject any emotional control over the music material, and places the emphasis on description and second-hand reporting, and the latter neglects the emotional input of those not skilled in musical performance. These approaches also do not take cultural influences into consideration, being concerned principally with the search for emotional cues in music generally.

2. The System - Specification Gathering

2.1 Initial phase - prototyping

The initial phase in creating the system was to move through a quick succession of prototypes, each gradually increasing in complexity, to see if the initial skeleton could support the proposed system. Once this was verified, then the system quickly evolved to the stage where it could perform basic musical operations on an inputted music file. This early stage of iterative design proved the plausibility of the technical solution proposed, and allowed the system to enter a more structured phase of development where musical, emotive and technical specifications were gathered and organized in a series of studies.

2.2 Transformations

The first of these was to locate and isolate suitable low-level musical transformations. This focused on the area of musicology, both general and cognitive, as both the general theories of music, and the experiences of others in the sphere of computer modeling, were of relevance. This study resulted in the formulation of a set of musical transformations, grouped in classes such as tempo, pitch and rhythm.

When considering emotional transformation in music, one has to consider the transformations themselves. They must be at the lowest level of musical control, the processes that alter the musical equivalent of punctuation, spelling and the use of italics in literature, for example. They cannot be of a high enough significance that personal preferences can be supported, or bias imparted. This does not mean that they need be powerless, however. Given a simple musical melody, it is possible to alter it significantly just by changing a few fundamentals; for example, speeding up the tune, changing the instrument playing it, changing the attack of the notes from long and smooth to short and jerky, and so on. By isolating and then grouping all these low level transformations into relevant sets, it should be possible to produce a toolkit, a set of building blocks, for altering the initial piece in any way required. The secret to the significance of the alteration would then be held in the sequence of simple steps, not in the steps themselves, somewhat like the use of a keyboard and a word processor to produce an emotionally-complex literary work despite only being able to add letters and punctuation.

Another important consideration relates to the types of transformations that will be used. The research presented here concerns itself with physically altering the structure of the input melody, of re-composing the piece via these transformations. It is not concerned with performance data, how someone plays a piece, but with the effect of altering a composition to change its emotive content. It does not attempt to map different inter-note intervals, for example, as was done by Bresin and Battel (Bresin & Battel 2000), where five pianists were analysed while playing the first 16 bars of the Andante movement of Mozart's Piano Sonata in G major K545. The research presented here is concerned with altering the piece as it is actually written, not how it is performed.

To initiate the process of compiling a list of transformations it was necessary to break music down to its fundamentals; from this research, a list of possible valid musical transformations was compiled. The final list is shown here:

Tempo, Pitch, Rhythm, Timbre, Harmony, Accompaniment, Dynamics, Drum Rhythm, Attack, Articulation, Scale.

2.3 Emotions

As the target for the system was the creation of emotional change, it was necessary to refer to research relating to human emotions, with a particular focus on what most psychologists refer to as the primary emotions, the basic emotions of which most emotional feeling is made up of. While there were a number of combinations available, the general consensus was towards a fairly well accepted set of emotions, and in the end it simply became a case of selecting the most suitable list, in terms of ease of data comparison, for inclusion in the experiments.

In the main study, it was this list of primary emotions that the users were asked to re-create through use of the system. For localisation purposes, this target list would probably have to be expanded, perhaps involving some aspects of how certain cultures perceive emotion in music, or maybe expanding the categories more into the area of moods; but the main thrust of the conversion would be in making the music, and its implied meaning, fit the locale better. In that respect, primary, or basic, emotions are an eminently suitable starting point given their generally accepted universality. For future work any arbitrary labels may be proposed for template generation, from emotions and moods through to descriptive media categories such as 'news', 'sports', 'up-tempo', 'youthful', and so on.

The list of emotions for the study was compiled after analysing the proposals of several psychologists; Robert Plutchik (Plutchik 1980, 2001), Shand (Shand 1914), Clynes (Clynes 1980), Izard (Izard 1991), Klaus Scherer (Scherer 1995), and Schopenhauer (Gale 1888). Of particular interest was Plutchik's classification system, which uses a three-dimensional circumplex model describing the relations among emotion concepts, which are similar to the colours on a colour wheel (Figure 1). The eight sectors are designed to indicate that there are eight primary emotion dimensions defined by the theory arranged as four pairs of opposites.

For this reason, Robert Plutchik's model was selected over those of Shand, Izard and Clynes, as it includes a pairing of opposing emotions, thus allowing further comparisons to be made between emotions as well as analyzing each emotion separately. His contention that more complex emotions are made up of combinations of the more basic ones is also of interest, as perhaps emotive combinations could lead to creating emotive musical content in a similar fashion, such as contempt arising out of a cross between anger and disgust. As a control, an option to model "no emotion" was also added. The list of emotions thus became:

Joy, Sadness, Anger, Fear, Acceptance, Disgust, Surprise, Anticipation and No Emotion.

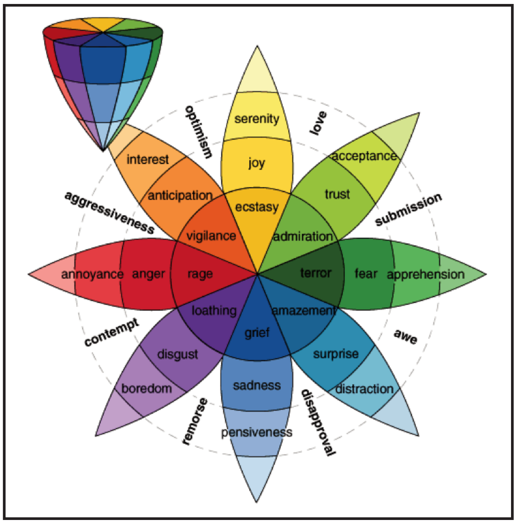


Figure 1 - Plutchik's three-dimensional circumplex model

3. The System - Design

At this stage the system had a requirements list for transformations, and a set of target emotions for users to attain, so work progressed in completing a working first version. Introspective testing, and informal peer reviews, ironed out any major functional issues in the system. A Preliminary Study was conducted, using the primary emotions selected, to help select suitably "unemotional" musical input pieces for use in further testing. This involved composing a list of new, and hence unknown, melodies for review, with those rated lowest in emotional content moving forward for use in later tests. A well known nursery rhyme was also included as a control.

3.1 Development

The development platform was X-Code and Interface Builder on an Apple Macintosh running OSX 10.4.x. Most of the music alteration functionality was written in ANSI C, with the user interface code being handled by Objective C. The system was developed from a small working framework in an iterative manner, using the spiral design methodology, with requirements being added as the system grew in size and capability. The advantage of this approach is that it is both incremental and iterative, and allows the development to start out small and benefit from enlightened trial and error throughout the development process.

The next phase of development was driven by informal peer reviews where colleagues tried out the system and highlighted any issues relating to either the user interface layout, or problems with the underlying functionality. From an architecture perspective, the system was now complete.

The final development phase involved users in a Pilot Study, which produced results that moved the focus on to improving usability, removing any remaining functional inconsistencies, and fine-tuning the pairing between the physical system and the tasks the participants would be called upon to compete.

3.2 Architecture

The overall system structure consists of three 'layers', named User Interface, Command and Realisation (Figure 3). One layer, the User Interface itself, is developed within Interface Builder, and is solely concerned with tracking, recognising and dealing with on-screen events. These events pass messages to the Command layer, which then fields these messages and decides what action to carry out, and what commands to pass on to the Realisation layer.

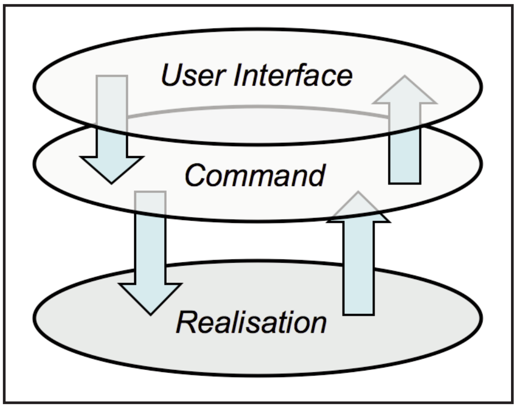


Figure 3 - System Architecture

The Command layer controls the application, listening to the interface for messages that signify button clicks or slider moves, and responds to these events by calling for the relevant action to be carried out. Some button presses require that the lower layer (Realisation) be called; some require management of the appearance of the user interface for controlling the focus of the user, and some deal with playing back sound files, or handling dialogue boxes for opening or saving script files, or inputting MIDI files.

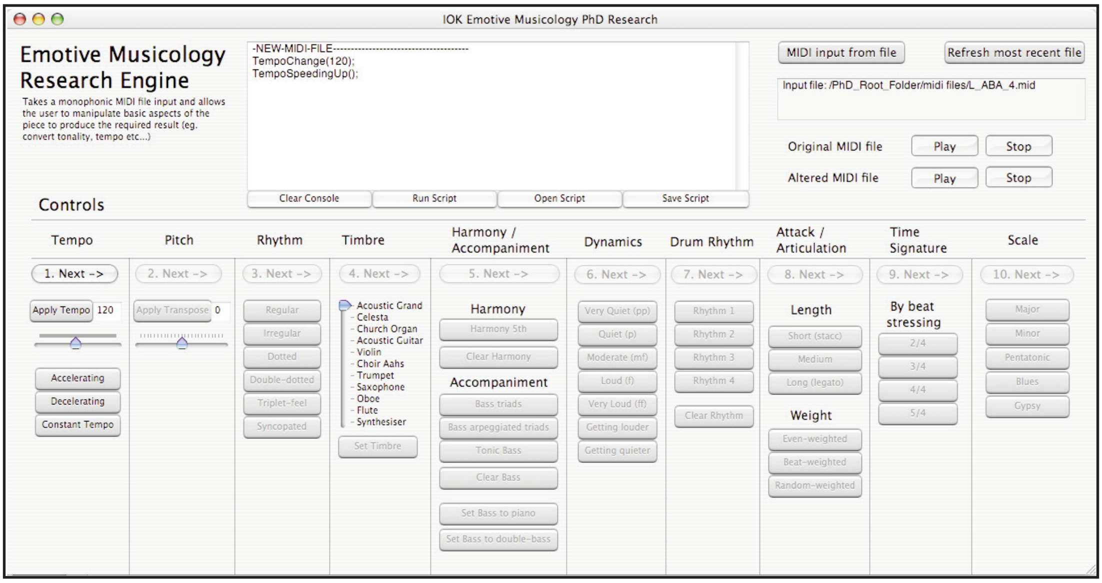
The Realisation layer does all the real work in the system. It controls the input and parsing of the source music file for alteration, storing the musical data into an internal array structure, and then applying whatever transformations the user has requested. It also performs its own housekeeping with respect to sorting, tracking track end points, adding MIDI track headers if absent in the inputted file, and quantisation to ensure all notes are correctly recognised and positioned as they would be in a musical score. Once all the work on the data has been completed it is outputted as a MIDI file for auditioning by the user. A special feature of the Realisation layer is its ability to deal with scripts in a form of batch processing as well as fielding single commands. This allows the user to submit a series of transformations in the form of a 'macro' to any inputted piece, and also means that the developer can create power functions in the Command layer that may use a series of two or more low level functions in combination.

4. The system

4.1 User Guide

The completed system (Figure 4) enables the user to perform low-level musical manipulations on an imported MIDI (Musical Instrument Digital Interface) file.

The first step involves selecting the MIDI file for transformation, via the "MIDI input from file"


IOK Emotive Musicology PhD Research

Emotive Musicology Research Engine

Takes a monophonic MIDI file input and allows the user to manipulate basic aspects of the piece to produce the required result (eg. convert tonality, tempo etc...)

```
-NEW-MIDI-FILE-----------------------------------
TempoChange(120);
TempoSpeedingUp();
```

Input file: /PhD_Root_Folder/midi files/L_ABA_4.mid

Apply Tempo: 120

Apply Transpose: 0

Figure 4 - The System

button. Once selected, the transformation controls become activated, in vertical columns of similar functions. These columns cover Tempo, Pitch, Rhythm, Timbre, Harmony, Accompaniment, Dynamics, Drum Rhythm, Attack, Articulation, Time Signature, and Scale. The Play and Stop buttons for the original and altered MIDI files remain active throughout the exercise so that the user is able to audition the altered piece after each transformation, and can review the original piece at any time for comparison.

When the user is happy with the emotion he has created for his locale or culture in the resulting piece, he saves the script created for that emotion using "Save Script", and clears the console prior to the next emotion via "Clear Console". Clicking on "Refresh most recent file" clears all transformations from the input file ready for the next pass.

4.2 Scripting

The key to the system's ability to collect emotive data is its scripting functionality. This allows the system to track every decision the user makes in altering the initial melody. However, while simply logging all user actions is a crucial requirement for the research, the ability of the system to re-read the scripts created, and to re-apply them to any new input file, gives it considerably more flexibility. It means that the system can, at a later date, enforce any user's preferences on any inputted melody, or that a generalized script for an emotion, created offline from data gathered using any text editor, can be imported and used separately.

4.3 Evaluation Methods

Analysis of the scripts produced by test subjects was done in an empirical manner using spreadsheets for totaling and averaging the data. The data was compared across emotions, to locate and isolate any noteworthy trends, and also across each participant, to isolate any trends created through participant favouritism rather than emotive effect.

Any scalar controls (tempo, transposition) were tracked in terms of direction as well as magnitude.

The aim of this analysis was to highlight any deviation from what could be regarded as random or normal, and also to compare the divergence between the different emotions, as this would show that the proposed approach had validity for capturing emotive content in music.

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Figure 5 - Chi Square Goodness of Fit formula

For the purpose of isolating any significant findings, the Chi Square Goodness of Fit test (Figure 5) was applied to the data, given its nominal/categorical nature. An alpha value of 0.05% was selected for the test for significance, although the sample size may be considered a little on the small side for robust findings at just nine participants. Even so, it gives a good indication of the trends in the data gathered. The two variables that displayed a parametric nature (tempo and pitch change) were analysed to calculate their mean and median values, and their standard deviations.

The collation process created nine tables of data for each participant, one for each emotion tested. These results were then totaled for each emotion, where individual button presses could be selected, or averaged by emotion where integer values could be selected, as for tempo and the pitch change. To isolate personal trends or favouritism by a particular test subject, the data was also totaled for each participant. For example, someone may simply like one option and select it for all emotions, and this data would show this trend and allow it to be accounted for in the overall data analysis.

Profile information was also compiled for each participant, via note taking during the study, informal discussion with the participant during and after the study, and via the questionnaire each participant completed after finishing the study.

Analysis of this profile information produced suggestions for the improvement of the system; and also data that could be cross-referenced with the empirical data, such as the level of musical expertise.

5. Results

5.1 Main Study

The results of the main study show that the idea of collecting emotive data from test subjects using low-level musical transformations definitely has merit, and some interesting trends are apparent even after analyzing the relatively small data sample gathered here.

Tempo is definitely a decisive factor, particularly in Joy, Sadness, Anger and Surprise. What is particularly interesting is that the average tempo for No Emotion is 105bpm, slightly slower than the input file at 120bpm. 120bpm is the default tempo for MIDI; hence its selection for the input file, but the study showed that this was regarded as having "too much" emotion. However, it should be noted that the No Emotion tempo sits quite centrally between the tempi selected for Joy and Sadness, showing its suitability as a median.

Some more examples of emotional opposites in the data, as suggested by Plutchik in his arrangement of the primary emotions as pairs of opposites, can be seen in Pitch between Joy and Sadness, and Anger and Fear; in Rhythm between Joy and Sadness, Acceptance and Disgust, and Surprise and Anticipation; in Dynamics for all emotional pairs; and also in many aspects of Attack Length and Articulation. Scale shows an almost bipolar split between Joy and Sadness, and also shows a lot of variance across the other emotions.

Moving the focus away from each transformation, and onto individual emotional templates, when they are compared in the combined chart (Figure 6), it can be seen that no one emotional template matches another, they are all unique. This demonstrates that each emotion is definitely mapping to its own set of preferences, and suggesting that there is validity in attempting to extract emotional content from music in this manner. It also demonstrates the strength of an entire set of transformations for an emotion, and hints that an emotional template may well be more than just the sum of its individual transformations.

It also tends to imply that a suggested possible result, namely the isolation of a unique combination of low-level musical transformations that could be regarded to represent some form of an emotional template, has been obtained. The close similarity in the results for individual participants for each emotion also lends support to the idea that these templates are demonstrating some consistency, and have some validity in representing their respective emotions. This therefore also implies that using these templates to transform other input pieces into a required emotional state is a viable proposition.

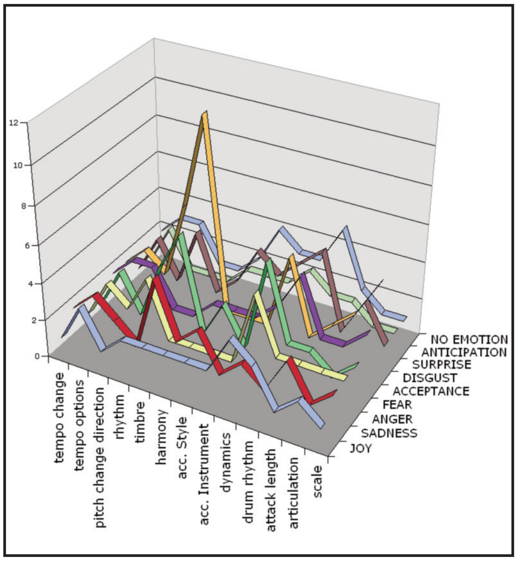


Figure 6 - Emotional Templates Compared

What is also noteworthy is the variance between skilled musicians and those with basic musical abilities in terms of the number of transformations auditioned. The skilled musicians took significantly fewer 'steps' to reach their required goal, but this disparity in the efficiency of the participants was not reflected in the outcome of the study, as it did not affect the close alignment in the data for each emotion. This suggests that the emotions recognised in the music by the participants are not strongly dependant on the level of their musical training, and that the emotional recognition process may be of a much more general cultural nature.

The data therefore suggests that the system has, firstly, been successful in capturing a series of results that differ enough from each other that many of them can be regarded as significant. Secondly, these results are distinct enough that they may be regarded as possible emotional templates. Taking these results, both empirical and observed, into consideration the research can be said to have produced a number of contributions:

- The possibility of extracting empirical data relating to human perception of emotions in music.

- The creation of a system to facilitate this analysis of the human perception of emotion in music, and to also facilitate the alteration of the emotional content of any inputted piece of music.

5.2 Follow-Up Study

The templates produced by the main study presented an excellent opportunity for verifying the data gathered by the system, simply by running the experiment in reverse. To this end, a study was created using a short piece altered emotionally by the system using averaged emotional templates gathered from the main study, and the test subjects were asked to categorise these pieces.

The study involved listening to the 9 pieces - each representing a different emotion - and categorising them by Emotion. No Emotion was included as a control, as before, and users were told that this category should represent an absence of emotion in the piece

As was the case for the main study, the table compiled by Plutchik defining the eight emotions was supplied to provide the test subject with the same definition of the required target emotions.

The results were split into two groups, those who were already familiar with the system (eight participants), and those who had never seen it (nine participants) and were judging the pieces purely on emotional content (Figures 7 & 8).

The results showed a significant outcome for the data collected for each emotion, although not all emotions have been correctly identified. What is also interesting is how closely the data matches between the two groups, suggesting that there is indeed a recognisable set of emotions in music, whether you are familiar with the system and its processes or just a listener to any emotionally altered piece. The strongest categorisation is for Sadness, with almost all test subjects rating it first. Disgust, Surprise, Joy and No Emotion are very high in certainty, followed by Acceptance and Anger, although some test

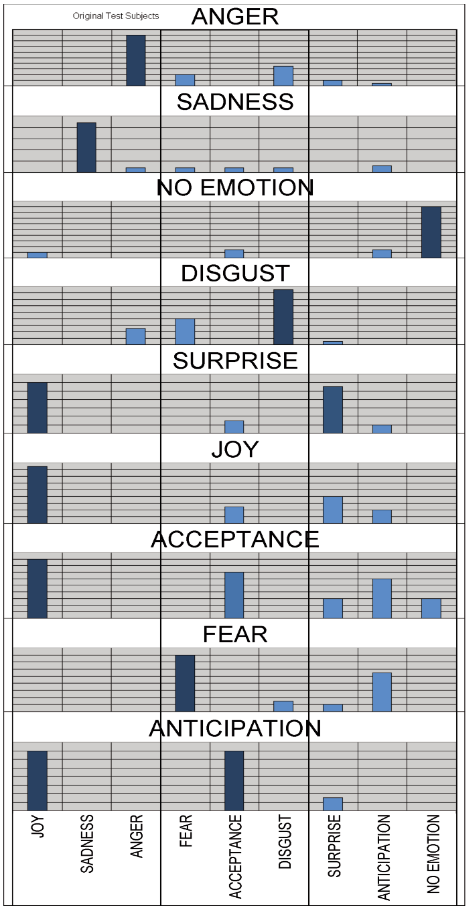


Figure 7 - Data for those familiar with the system

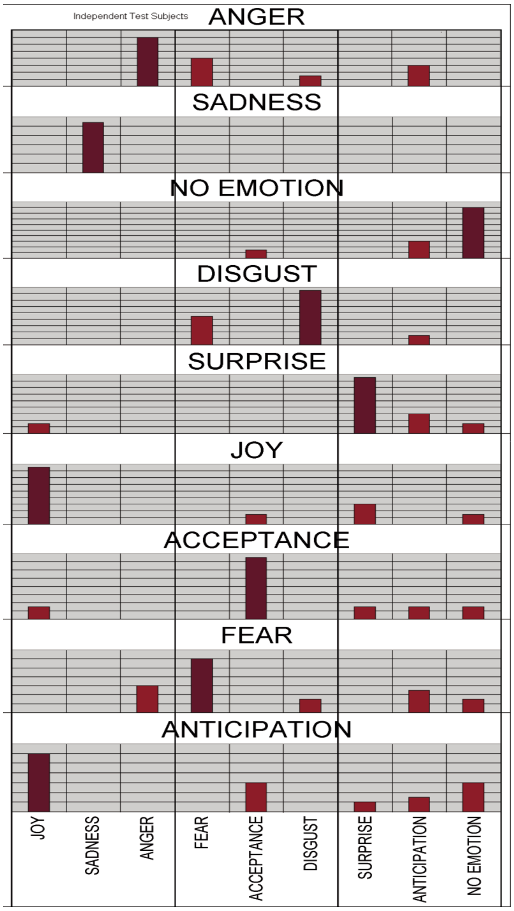


Figure 8 - Data for the independent test subjects

subjects also linked the Joy piece to Surprise. Fear also produced significant results, although it was sometimes confused with Anticipation. Acceptance was categorized as Joy by the group familiar with the system; in contrast, the independent group classified Acceptance correctly. Anticipation caused widespread confusion, receiving no votes at all from those familiar with the system, and just one vote from the independent test group. Most test subjects categorised it under Joy, with Acceptance and No Emotion being the next most favoured.

This study essentially closed the loop on the emotional research undertaken, feeding the results back for verification, and demonstrated that pieces generated via the extrapolated emotional templates were, for the most part, correctly identified by a listening test that can be viewed as independent from the Main Study. Some emotions were classified more strongly than others, but where an emotion was correctly identified, statistical calculations suggest that this result should be regarded as significant. The small sample size needs to be taken into account, however, but the early signs are promising for the approach used in this research.

6. Discussion

The results gathered by this study show much promise for an interactive approach to music analysis and alteration. The benefits are observable; clear, empirical data captured from precise on-screen user decisions, avoidance of confusion through bias, introspection, and descriptive issues, ease of usability across a broad musical ability spectrum, and easy capture of emotive templates for re-use. The follow-up study also demonstrated that listeners were able to correctly identify the intended emotion or mood in pieces of music that had been altered by the templates gathered by the system.

To summarise the capabilities the approach provides:

- The ability to capture a user's preferences, in the form of a template, in response to a request to induce an arbitrary mood, emotion or categorization in any piece of music.
- The ability to store this template for re-use on other musical material as required.
- The ability to generalize the data gathered across a number of users for any specific template descriptor; for example, sadness.
- This then implies the ability to segment any generalized findings into any required demographic, such as locale, culture, gender, or whatever.

This, of course, is a general overview of the possible capabilities of the research presented here. What needs to be done next is to consider the possible applications of this functionality.

6.1 Applications for Localisation

Now that the concept of capturing emotive templates for music has been demonstrated, the next phase is to ask how this system could be leveraged in a music localisation environment.

Firstly, from a standards viewpoint, it would be necessary to specify that all symbolic digital musical content for musical localisation is represented in MIDI format, at least until a better format is adopted. Of note are the recent developments in MPEG-4, as it is suggested (Bellini, Nesi & Zoia 2005) that the integration of SMR (symbolic music representation) in the MPEG-4 standard will enable the development of many new applications. Broadening the base of the musical representation would allow more space for handling multimedia opportunities, such as guitar tablature, libretti for operas, and audio-visual components. As of the time of writing, standardization is still in progress. The MPEG AHG (adhoc group) is pursuing this standardization under the auspices of the MUSICNETWORK, a group funded by the European Commission to help bring music into the interactive multimedia era.

Secondly, a set of categories for the templates would need to be proposed, such as: Basic Emotions (Happy, Sad, Angry, Fearful and so on), or Media Categories (Corporate, News, Sport, Youth Culture, Up-Tempo, and other such terms used within the media industry). The categories and sub-categories would then be used to tag the relevant music content so as to facilitate the localisation process. These categories could always be expanded later as required.

From the perspective of demonstrating the on-demand localisation of existing online music, initial user trials could involve using the existing system to create a number of culturally suitable versions of the musical content in advance, with the website then selecting the correct match to the user's profile when they connect to the page in question. If a culture was not available, default fallback behaviour could be used, as is currently done for language selection in .NET localisation.

The finished version, however, would be much more sophisticated, having the music transformation routines held on the website servers (possibly in Java for platform independence - as MIDI files, being a symbolic store of performance data, take up very little space and do not place a large load on available processing power), and the specific cultural versions of the required categories and sub-category templates being held on the user's machine as part of their profile in a similar manner to the way in which fonts are currently handled. The web server would then use these user templates to select the transformation functions that matched up to the category tag linked to the musical data, and output a culturally-localised music stream.

Possibly the biggest area of work would relate to the creation of the templates themselves in all the required categories, sub-categories and cultures. There are two main routes to pursue here, either commissioning studies to create musical templates for all required locales or cultures, or by enabling users to create their own templates, with the opportunity to upload these new templates so others can use them also.

The former approach would be similar to commercial localisation practice, would cost money, and would probably lead to a similar divide between commercially viable locales and those that are not regarded as being as lucrative.

The latter could be regarded as being a variant of the crowd-sourcing model, where users would be able to create their own cultural templates for music, possibly with the option of uploading these templates for use by the global community. Then online peer voting could establish the templates viewed as most suitable by the listening community.

The term 'Crowdsourcing' was first coined by Jeff Howe (2006) and later defined as "the act of taking a job traditionally performed by a designated agent (usually an employee) and outsourcing it to an undefined, generally large group of people in the form of an open call." (Howe 2009). Howe also refers to it as "the future of corporate R&D", citing the example of InnoCentive, the "research world's version of iStockPhoto" (Howe 2006). What is of more interest here, though, is crowdsourcing motivated simply through personal desire to make a contribution, such as demonstrated by contributors to the translation of Facebook. Their reward is recognition from their peers, and perhaps personal satisfaction. No money changes hands.

What is proposed here is the setting up of a simplified, web-based version of the music modification application that would be open to anyone who wishes to contribute, and create templates for their own particular locale or culture, particularly if they feel the existing templates for emotions, moods or categories do not accurately reflect their beliefs after auditioning music modified by them. In fact, it would be hoped that this form of 'national pride' would provide the motivation for contribution, as it has done for some minority languages on Facebook. Quality enforcement, and the avoidance of online 'vandalism', would be realised by including a peer voting system similar to that implemented by Threadless, the web-based t-shirt company (Brabham 2008). This would allow visitors to the site to vote on existing templates.

While there would be an initial set of categories and sub-categories set up (as proposed earlier), it would be wise to allow contributors to also suggest other categories, thus ensuring nothing obvious had been missed. A list of nationalities and languages would also have to be provided, to delineate the templates inputted by culture or locale. Once the initial data was gathered, targeting a much larger user group than in the studies described in this paper, then the true relevance of music localisation could be assessed, as well as the level of demand for such a service.

Further development could see the implementation of these templates in actual websites, and perhaps the facility for the personalisation of the templates on each user's local machine, bringing localisation to possibly its ultimate conclusion, a "locale of one" (Wade 2009), where each user has their own preferences on their own computer for the affective content of all musical content they hear, although the strength of the link to the localisation locale or culture in this instance could be called into question.

In conclusion, this research demonstrates the plausibility of creating a system to capture cultural-specific templates in music by involving the participant directly in the process of creating those templates. While there is still a lot of work to be done from a propagation standpoint, this work presents a new research possibility in the fields of cognitive musicology and localisation - the possibility to perform cultural modification on music for the localisation of online digital musical content - and the system constructed is a strong foundation for the further development of such applications.

Acknowledgements

The author would like to thank his PhD supervisor for her support and advice, and also Enterprise Ireland and the Interaction Design Centre in the University of Limerick for supporting the initial research. This research is also supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) in the Localisation Research Centre in the University of Limerick.

References

Bellini, P., P. Nesi and G. Zoia, 2005. Symbolic Music Representation in MPEG. IEEE
MultiMedia 2005, Vol. 12, Issue 4: 42 - 49.

Brabham, D.C., 2008. Crowdsourcing as a Model for Problem Solving: An Introduction and Cases. Convergence 2008; 14; 75

Bresin, R. and G. U. Battel, 2000. Articulation Strategies in Expressive Piano Performance: Analysis of Legato, Staccato, and Repeated Notes in Performances of the Andante Movement of Mozart's Sonata in G major (K. 545). Journal of New Music Research, 29: 211 - 24.

Clynes, M., 1980. The communication of emotion: theory of sentics. Theories of Emotion Vol. 1, R. Plutchik and H. Kellerman (eds), Academic Press, New York, 171 - 216.

Fact Monster, 2004. What Colors Mean. Pearson Education, Inc. 15 June 2004.
http://www.factmonster.com/ipka/A0769383.html

Gale, H., 1888. Schopenhauer's Metaphysics of Music. New Englander and Yale Review, Volume 48 Issue CCXVIII, 362-368.

Good M, B L Picot, S G Salem, C Chin, S F Picot & D Lane, 2000. Cultural Differences in Music Chosen for Pain Relief. Journal of Holistic Nursing, Vol. 18, No. 3, September 2000 245-260

Gregory, A. H., N Varney, 1996. Cross-Cultural comparisons in the affective response to music, Psychology of Music, 24, 47-52

Howe, J., 2006. 'The Rise of Crowdsourcing', Wired, Issue 14.06.

Howe, J., Crowdsourcing: A Definition. Available at:
http://crowdsourcing.typepad.com/
(Accessed November 26, 2009).

Izard, C. E., 1991. The psychology of emotions. New York: Plenum.

Jacko, J.A., 2009. Human-Computer Interaction: Design Issues, Solutions, and Applications. CRC Press, 2009

Kugel, P., 1992. Beyond Computational Musicology. Understanding music with AI: Perspectives on music cognition, MIT Press, 30 - 47.

Laske, O. E., 1992. Artificial Intelligence and Music: A cornerstone of Cognitive Musicology. Understanding music with AI: Perspectives on music cognition, MIT Press, 3 - 28.

Meyer, L. B., 1956. Emotion and Meaning in Music. Chicago: Chicago University Press.

Morrison, S.J., Demorest, S.M., Aylward, E.H., Cramer, S.C., Maravilla, K.R., 2003. fMRI investigation of cross-cultural music comprehension. NeuroImage 20, 378-384.

Nemzow, M., 2001. Juggling Global Monetary Complexities. Volume 12 Issue 1 of MultiLingual Computing & Technology, Sandpoint, Idaho, USA.

Plutchik, R., 1980. A general psychoevolutionary theory of emotion. Theories of Emotion Vol. 1, R. Plutchik and H. Kellerman (eds), Academic Press, New York 171 - 216.

Plutchik, R., 2001. The Nature of Emotions. American Scientist, 89: 344 - 350. July/August.

Scherer, K. R., 1995. Expression of Emotion in Voice and Music. Journal of Voice, 9(3), 235-248.

Shand, A. F., 1914. The Foundations Of Character - Being a study of the Emotions and Sentiments. Macmillan and Co. Limited.

Wade, V., 2009. Supporting a Locale of One. LRC XIV "Localisation in The Cloud": The 14th Annual Internationalisation and Localisation Conference, Limerick, Ireland.

Walker, R., 1996. Open peer commentary: Can we understand the music of another culture? Psychology of Music, 24, 103-130

Micro Crowdsourcing: A new Model for Software Localisation

Chris Exton[1], Asanka Wasala[2], Jim Buckley[1], Reinhard Schäler[1],[2]
[1]Centre for Next Generation Localisation,
[2]Localisation Research Centre,
University of Limerick
www.cngl.ie
www.localisation.ie
chris.exton@ul.ie; asanka.wasala@ul.ie; jim.buckley@ul.ie; reinhard.schaler@ul.ie

Abstract

One obvious flaw in the concept of the knowledge society is our collective failure to date to provide equal access to information and knowledge across languages. We are a long way away from the ideal world, where, as Muhammad Yunus, winner of the 2006 Nobel Peace Prize said, there would only be one language in the information technology (IT) world - your own (Yunus 2007). While the US$16b mainstream localisation industry likes to see itself as the vehicle that is removing this barrier to universal access to digital knowledge and information (i.e. language), in reality it is making limited impact on the widening gap between the information rich and the information poor.

Crowdsourcing has been described as an approach to address the shortcomings of current mainstream localisation, allowing the localisation decision to be shifted from large corporations to service users, thus making IT available in more languages. This paper proposes and describes a new model of crowdsourcing which may provide a platform by which the "equal access to information and knowledge" might be achieved.

Keywords: *localisation, digital divide, micro crowdsourcing, real time localisation*

Introduction

Current software localisation efforts are largely driven by the economic imperative of short-term return on investment. The localisation decision, i.e. the decision on the languages and locales to be covered by the mainstream localisation effort, is almost exclusively determined by the size of the market a language or a locale represents. Therefore, software is translated for the approximately five million speakers of Danish, but not for the 27 million speakers of Amharic, the national language of Ethiopia. However, there is also a long-term return on investment issue to be considered for commercial organizations. Digital publishers have recognised that, without pursuing deployment in large but currently unviable locales, their product will have limited exposure that may jeopardise future, larger market gains when those locales develop economically.

In addition, the social effects of this short-term economic imperative are grave. Access to information technology is restricted to those speaking the languages of the global north while excluding those speaking the languages of the global south. The majority of people living on this planet cannot share their knowledge in the digital world, nor do they have access to existing knowledge. Organisations engaged in localisation activities not primarily for commercial but for social, cultural, political or developmental reasons, see crowdsourcing as a mechanism to connect with their communities, through reduced cost.

The concept of crowdsourcing was first described by Jeff Howe in his now famous article in Wired Magazine in 2006 (Howe 2006). He described it as the harnessing of a community/group of people to perform a task traditionally undertaken by employees. Crowdsourcing has been taken up enthusiastically and has resulted in almost 8.2 million hits in Google's search engine. It has featured as a major topic at recent, seminal localisation-industry events, such as Localization World 2009 and LRC XIII. This is because both commercial and altruistic organizations see it as a mechanism to lower the cost of localisation, enabling them to enter currently inaccessible locales (Rickard 2009). In addition, (Losse 2008) in her keynote at the LRC conference, stated that organizations like Facebook pursued crowdsourcing because it also produced higher quality translations.

Consequently, attempts by companies such as Microsoft, Facebook and Google to create crowdsourcing frameworks that allow volunteer translators and localisers to translate digital content into marginally commercial languages are seen by the industry as having delivered very promising results.

While commentators seem to agree that the main issues around crowdsourcing in the localisation space are control, quality and motivation, there is still a lack of comprehensive studies on any of these issues. In addition, another central issue of crowdsourced localisation, i.e. the need for the localisation decision to be shifted from multinational corporations to the user (Howe 2006), has only been raised on occasion and not as a central pre-requisite for the success of any crowdsourced localisation effort. (Schäler 2008).

Based on these issues, this paper considers how the practices and experience from the open source and Web 2.0 communities could provide a path for software localisation to make Muhammad Yunus' dream a reality.

1.1 The Cathedral and the Bazaar

The "Cathedral and the Bazaar" is an essay by Eric S. Raymond on software engineering methods (Raymond 1999), based on his observations of the Linux kernel development process and his experiences managing an open source project. In it he describes how the traditional software development paradigm could be viewed as hierarchical and tightly planned; Raymond likens this view to a Cathedral which is monolithic and obviously architected by some controlling authority. The open source development paradigm however, he continues, could better be likened to a market or bazaar, where it is obvious that industry of some kind is occurring but there seems to be no or little central authority or control.

Traditionally, mainstream software localisation has been tightly controlled by multinational corporations. They strictly managed everything from the localisation decision itself to the selection of an appropriate localisation process, the use of certain terminology and translation memories, and the deployment of adequate tools and technologies.

This model is not unlike the Cathedral model described by Raymond, with its central control and tight management through a number of levels of activity and quality control, to ensure a suitable and tested final product. Indeed, the type of software often used to support and control the activity of localisation, (essentially customised project-management systems such as Idiom Worldserver) provides strong evidence of the approach chosen by mainstream localisation. However, such a 'Cathedral' model brings with it the implicit need for large coordination efforts and subsequently high costs.

The 'Bazaar' model, in turn, is associated with the open source community, and requires lesser controlling authority. Work is somehow carried out in an almost chaotic, community-driven manner. This has provided the business and technical communities with a suite of software, including Linux (Raymond 2001), Apache (Mockus et. al 2000) and Openoffice (Feller and Fitzgerald 2002), upon which many companies rely heavily today. These systems are proof that a community-driven, open source model can also deliver quality software systems.

Indeed the open source community model may provide a paradigm to address some of the problems faced by the localisation community in relation to their desire to expand into underserved markets and to break down the digital divide. However, many initiatives addressing underserved markets today, such as the ones initiated by Facebook, Microsoft or Symantec, still have a central authority driving the localisation effort, rather than being bottom-up and community-driven.

This paper proposes an alternative approach to the idea of crowdsourcing in relation to the translation of software systems. In this model, individual users translate elements of a system and its documentation as they use them in return for free access to these artifacts. Periodically, the elements of the system and documentation translated by the individual translators are gathered centrally and aggregated into an integral translation of all, or parts of, the system.

2. Approach to Micro-Crowdsourcing

Consider a software package such as Open Office that has been developed for a purely English speaking audience. Even if this product were designed to facilitate its easy adaptation into other languages it would still require the effort of either a number of altruistic individuals or the coordinated effort of expensive professional localisers to make this product available in another language.

Where there is no/limited immediate economic imperative for the digital content publishers, such as in the case of open source software or voluntary organisations aiming to bridge the digital divide, one solution might be the automatic translation of content into non-commercially viable languages. Although this option might be preferable to simply ignoring these languages, automatic translation is not yet at a stage where such a product could be released with confidence.

Imagine however, developing a software application such as Open Office that allowed a community of users to update the user interface in situ, either directly from the original English version or perhaps working from a less than perfect machine translation. The update could be enabled via a simple popup micro localisation editior that would allow them to change UI text in situ simply by ctrl-clicking on any text that is displayed.

This editor may have to enforce constraints on the translation, such as restricting string length, and could perhaps include appropriate translation memories and standards to assist in the translation. As a ctrl-click could be applied to any displayable text area, error messages and help information messages could also be included as translatable material.Indeed, the editor may even go as far as allowing graphical replacement of certain artefacts.

The result would be a set of textual (and possibly graphical updates for each user. Then suppose that each update-set could be automatically gathered in a central repository that would, in turn, push update events back to the community of users, periodically or on-demand. This would update their product with the latest translations. Imagine that these users could, in turn, quality assure the updates and re-instigate the cycle, in the same way that Web 2.0 communities like Wikipedia reach consensus by iterative refinement.

Such an approach would represent a radically novel approach to localisation requiring a novel architecture, as demonstrated in figure 1. On the far left, we see a central server that receives and sends updates to and from individual deployments of a

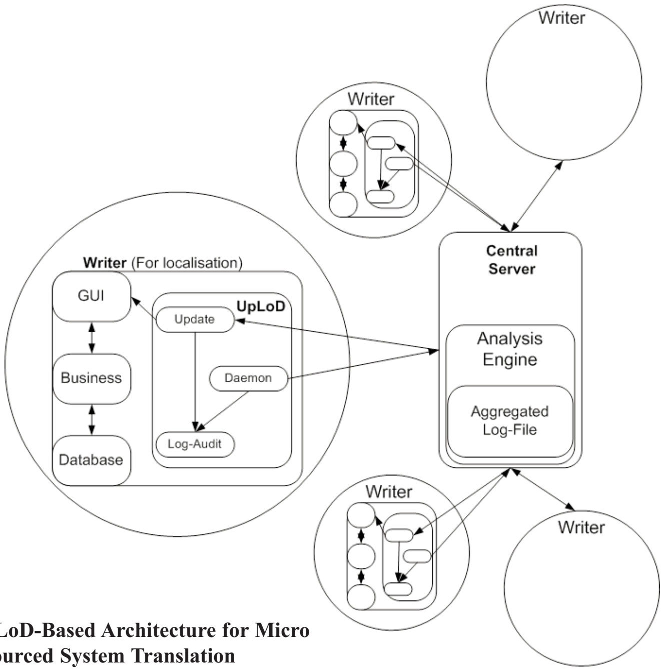


Figure 1: The UpLoD-Based Architecture for Micro Crowd-Sourced System Translation

software package called 'Writer', one of which (to the left of the figure) is substantially expanded.

As can be seen from the 'expanded Writer', the three tiers of the application are augmented by an 'Update-Log-Daemon' (UpLoD) module. This UpLoD module allows the user to update the user interfaces as they use the system and log the changes in a local audit file. The records in this local audit file contain unique identifiers for the GUI elements that have been changed. The identifiers are associated with the pre-translation and post-translation. Periodically, a Daemon trawls the audit log and, on finding new records, passes updates to the central repository on the server.

These updates can be handled in a number of ways. For example in publically edited wikis, revision control enables a human editor to reverse a change to its previous version. For a "Micro Crowdsourcing" system it is possible to consider that there might be a limited number of trusted editors (self moderating) for a specific language group to tidy up the localisation in this fashion. A version control system would then enable editors to build a release package on a periodic basis based on the influx of micro changes from standard users. This would serve to drastically decrease the number of changes and updates to the UI and avoid updates with new translations on an ongoing basis.

For a more automated approach the server might periodically analyse the update set of all users, based on an aggregate consensus, and may be able to recommend the changes to be made to the other versions of 'Writer'. These changes are captured by the UpLoD module and update the GUI correspondingly.

Another open source development concept which could be adapted to suit the "Micro Crowdsourcing" model is that of distributed revision control. Distributed revision control is built on a peer-to-peer approach, unlike the centralised client-server approach classically used by software versioning systems such as CVS. In a distributed revision control system each peer maintains a complete working copy of the codebase. Synchronization is conducted by exchanging patches (change-sets) from peer to peer, a more in-depth discussion of the process is described by Noah and Adam (2009).

Regardless of the version control system that will be used, the translation is carried out in an incremental, ad-hoc manner by a community of (not necessarily experienced) "translators", each of whom would double as a proof reader for each other's work.

Once we allow all registered end users to become translators or localisers, we spread the workload over a large user base. The limiting factors would be the number of bi-lingual speakers with access to computers and internet connectivity. However, even this limiting factor could be overcome by offering monoglot users, familiar with the software, access to suitable translation aids including machine translation, translation memories and terminology databases.

To a large degree, a similar model already exists in the Wikipedia community where content may be added and amended by any registered user. Quality and precision, issues discussed as highly problematic in the context of crowdsourced localisation, are in this case simply promoted by the fact that any reader of Wikipedia can register and thereby correct or update any particular entry. This phenomenon can be likened to the "many eyes" principle associated with open source. This phrase was coined by Linus Torvalds (Raymond 1999) states "Given a large enough beta-tester and co-developer base, almost every problem will be characterized quickly and the fix will be obvious to someone." It simply describes the notion that, since open source code can be viewed and potentially changed by anyone who cares to look at it, the number of bugs that are caught and fixed increases dramatically compared to proprietary driven development. Likewise, it is envisaged that this "many-eyes" characteristic of the UpLoD architecture will promote an increasingly stable, high quality, and locale-specific application over time as users are empowered to become part of the localisation process.

In addition, the software is translated for free by volunteers, provided that the digital publisher is willing to deploy its un-translated or automatically pre-translated version to registered volunteers in each locale. In open source scenarios and scenarios where the aim is to bridge the digital divide, this will not be a concern. However, even in commercial scenarios the corporation may consider the exposure gained from having a localised version to be worth the loss of potential license revenue from registered volunteers. This would be particularly true in the case of emerging markets where the possibility of sales might be low at the moment, but where early exposure to localised systems could lead to commercial opportunities later.

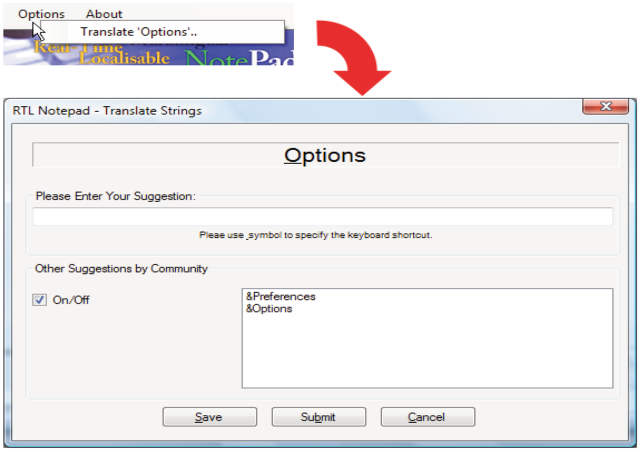

Options About
Translate 'Options'..
Real Time Localisable NotePad
RTL Notepad - Translate Strings
Options
Please Enter Your Suggestion:
Pleae use _symbol to specify the keyboard shortcut.
Other Suggestions by Community
☑ On/Off
&Preferences
&Options
Save
Submit
Cancel

Figure 2: The Localisation Dialog of the RTL Notepad application

3. Proof of Concept Implementation

A proof of concept prototype of this architecture was created to validate and refine this approach. The prototype consists of two components: the central server component and a simple RTL (Real Time Logging) Notepad application which imitates the "Writer" of Figure 1. The UpLoD module was implemented and integrated in the RTL Notepad application in addition to its generic text editing functions. Due to its simplicity and portability the Portable Object (PO) file format was chosen as the format for the local audit file.

In the RTL Notepad, simply right-clicking on text inside any UI element brings up a context menu where users can enter into a 'localisation dialog' (See figure 2). From this window, a user can translate the selected UI strings. The changes are reflected in the UI in real-time. Options have been provided to users in the localisation dialog for the online transmission of their changes to the central server or for the offline saving of the translation to the local audit file for later batch transmission to the server.

In this prototype, we propose an automated translation voting mechanism to ensure the quality of the translation. For this purpose, the server maintains a database containing translations and the number of votes for each translation (ie: the number of users who have suggested that translation), for each language.

In the following sections, the main phases of the localisation process associated with this architecture are explained in more detail. For illustration purposes, screenshots of the RTL Notepad in English and its translated version in Sinhala are given in Figure 3 and Figure 4.

3.1 Initialisation

The RTL Notepad application can be configured to update its UI by connecting to the central server or by reading from its local audit file, i.e. in line with the changes made by its immediate user. If the Notepad is configured to connect to the central server, it will retrieve the translations from the server and will update its user interface accordingly. This may result in overwriting the customisations already made by the user and they should be alerted to this possibility before they agree to the update or they should be alerted on a (UI) string-by-string basis.

In the first scenario, i.e. when the RTL Notepad is configured to update its UI using the information obtained from the server, the RTL Notepad will send an HTTP request to the central server stating the user's language, as configured in the RTL Notepad application. Then, the server will send an XML response which contains a list of source-target translation units for all the UI elements. This process is illustrated in figure 5. The server will generate the XML by choosing the translation with the highest number of votes, for each UI string. The RTL

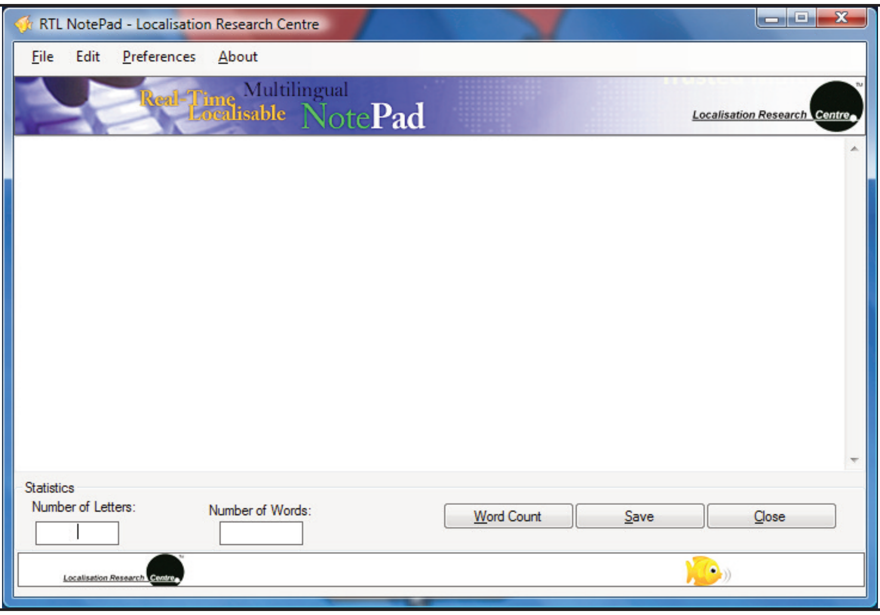


Figure 3: The RTL Notepad application in English

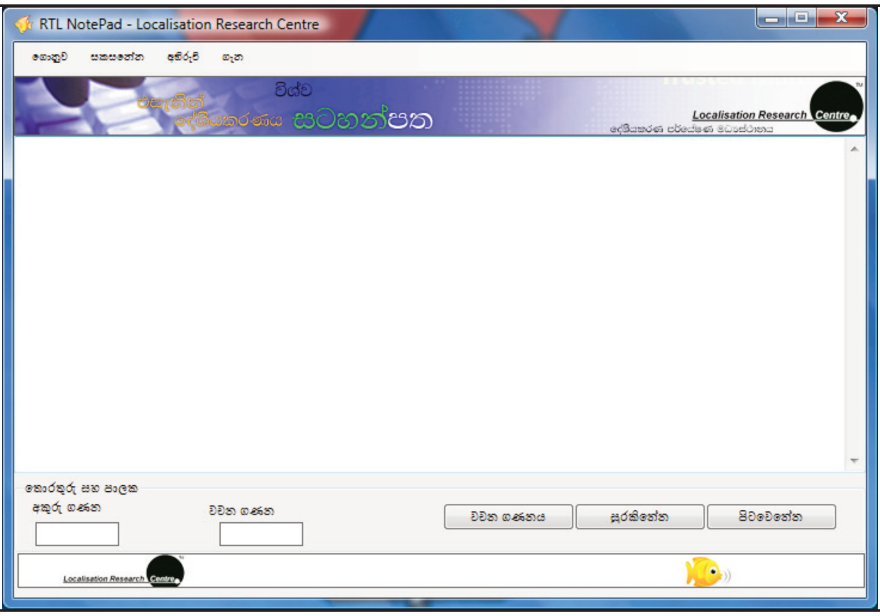

RTL NotePad - Localisation Research Centre
ගොනුව
සකසන්න
අභිරුචි
ගැන
විශ්ව
එසැනින්
දේශීයකරණය සටහන්පත
Localisation Research Centre
දේශීයකරණ පර්යේෂණ මධ්‍යස්ථානය
තොරතුරු සහ පාලක
අකුරු ගණන
වචන ගණන
වචන ගණනය
සුරකින්න
පිටවෙන්න
Localisation Research Centre

Figure 4: The RTL Notepad application in Sinhala

Notepad will process the XML response and update its UI. An illustrative XML response is given in figure 6 (where the GUID tag uniquely identifies the UI element).

There will be situations where different translations for the same UI string have the same number of votes. To handle such potentially 'thrashing-like' scenarios, the RTL Notepad will show a special dialog, allowing users to choose their preferred translation during start-up. In order to minimise such translation conflicts in the future, the user preferences are sent back to the server so that the server will increase the votes for the relevant translations. They will also be stored locally so that the user does not have the repeat his/her choice.

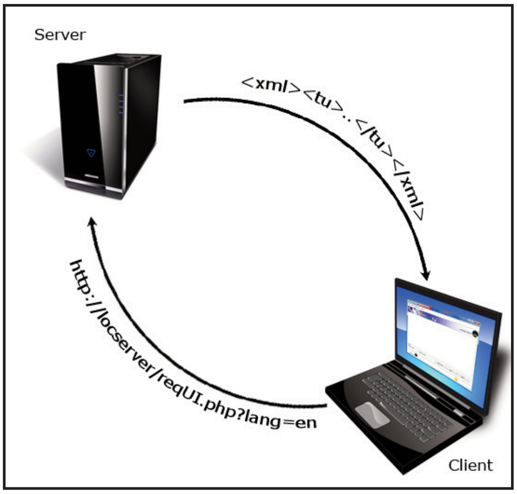
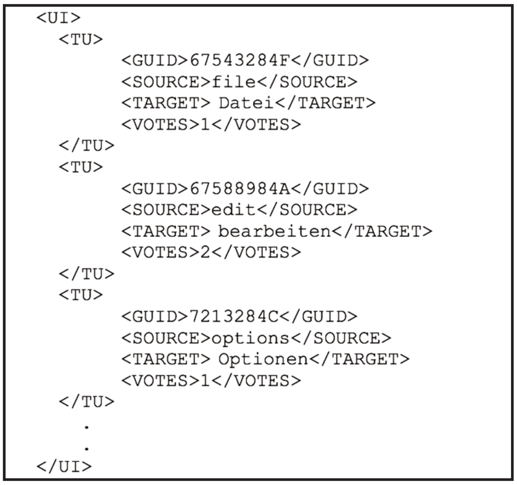


Figure 5: UI String Translations Request and Response Process



```
<UI>
  <TU>
        <GUID>67543284F</GUID>
        <SOURCE>file</SOURCE>
        <TARGET> Datei</TARGET>
        <VOTES>1</VOTES>
  </TU>
  <TU>
        <GUID>67588984A</GUID>
        <SOURCE>edit</SOURCE>
        <TARGET> bearbeiten</TARGET>
        <VOTES>2</VOTES>
  </TU>
  <TU>
        <GUID>7213284C</GUID>
        <SOURCE>options</SOURCE>
        <TARGET> Optionen</TARGET>
        <VOTES>1</VOTES>
  </TU>
     .
     .
</UI>
```

Figure 6: Typical Server XML Response

3.2 Translation Submission Process

Users can submit translations of UI strings to the central server through the localisation dialog of the RTL Notepad. User submissions will be directed to the central server as HTTP requests. Upon submission, the central server will query its database to see whether the submitted translation already exists. If so, the server will increase its votes. Otherwise, the translation will be added to the relevant language table, initialising its number of votes to one.

3.3 Community Suggestions Retrieval Process

In the localisation dialog of RTL Notepad, an option is given to retrieve the suggestions of the user community. Once the 'View other suggestions by community' option is selected in the RTL Notepad, it will send an HTTP request to the server asking for community translations for the selected UI string. The server will send an XML response to the client RTL Notepad application containing all the suggestions for the given UI string for a given language. The UpLoD module in RTL Notepad will process this XML and list these suggestions in its localisation dialog, in the order of number of votes received, as illustrated in figure 2. The users then can choose their preferred suggestion to be used in the GUI of that version of RTL Notepad. Once a user selects a suggestion, the suggestion will be sent back to the central server as an HTTP request. The server will then increase the votes of the given suggestion.

4. Outstanding Challenges

Of course, localising software is not as simple as portrayed in this prototype. Not only does text have to be changed: holding boxes have to be resized, and images may have to be replaced, for example. However if tools that facilitate localisation were incorporated into an UpLoD-type architecture, it would not be unreasonable to expect that the need for these changes could be covered satisfactorily. After all, such changes are currently performed in existing localisation efforts. The model proposed here suggests piggy-backing such functionality into the "Update" component of the system deployed.

There is also the possibility that volunteer translators would focus their efforts on only a small proportion of the user interface. This proposition is based on Pareto's Principle (Bookstein 1990) which, to paraphrase for this context, suggests that most users of large applications will only use a small proportion of its functionality. If translators choose to translate as they use, or choose to do the translations that others will see, rather than translating holistically, it is likely that translation coverage will be patchy and will result in a 'pidgin' system made up of translated 'frequently-used' facilities and untranslated 'infrequently-used' facilities. This may prove sufficient for the majority of users, but runs the risk of frustrating users who have more demanding requirements. However, frustration can become a motivating factor if the user is empowered to subsequently change the associated UI strings.

Another potential challenge is that the voting mechanism proposed may prove insufficient and ineffective; specifically, there is the possibility of 'thrashing', where two individual translators, or groups of translators, have very strong and conflicting ideas about the translation required for specific GUI elements. In such instances, the 'Analysis Engine' of the central server would need to intervene, analysing the central logs, deriving the appropriate translation, possibly with human intervention, and locking future changes.

Indeed, we see this 'thrashing' problem as being one of the main issues with this approach. Imagine, as a user, you customise the interface and then send your changes to the server. Imagine then, retrieving the server-side customisations and finding that very few of your changes had survived. This is a micro-form of thrashing that would probably be prevalent, particularly if there were a wide number of users customising the interface - a measure of the approach's success. Such negative feedback might discourage the user from making further changes to the interface and result in a fall off in localisation activity over time. Indeed, it might discourage them from using the application itself, as the interface they strove to create has been destroyed by the server-side customisations. Hence, as mentioned in section 3.1, we see a strong role for 'change alerts' and the option to opt out of server-side customisations as core for the users of this approach

The voting mechanism currently implemented in the prototype takes no account of user quality, an attribute that could easily be calculated from the available data (a simple measure could be the percentage of each user's suggestions that equate to the customisations with the highest vote). This additional information could be used to resolve ties, where equal numbers of votes were obtained for two or more different translations, to resolve thrashing or, more generally, as a weighting on the votes.

It may be that user submissions might have to be reviewed by human experts (preferably by a pool of linguists) prior to committing to the server's database. This additional step would ensure the quality of the translations to be used in UI elements in terms of criteria such as relevancy, accuracy, suitability, and consistency. However larger scale deployments, where a bigger community is involved, may well counteract this potential issue.

It is noteworthy to mention the programming difficulties that may be encountered when developing UpLoD-architecture-based applications. The development of the prototype revealed that some UI widgets and built-in UI components such as file open dialog, printer dialog etc. provided by several programming languages are tightly integrated with the underlying platform and hence cannot be modified to incorporate real-time localisable features.

Notwithstanding, UpLoD-architecture-based applications are easy to develop using programming languages that support object oriented programming (OOP), especially if their UI development components are loosely coupled with the operating systems and the rest of the system's components. Indeed, ideally, the "real-time localisability" should find support within the operating systems and the programming languages themselves.

However, as long as this is not the case, it has to be acknowledged that there is an overhead associated with the UpLoD architecture that adds to the expense of this initial development, so ideally this architecture should be as pluggable-and-playable in nature as possible, and this is seen as an area of future research for the group.

This overhead may be increased in a large-scale UpLoD system. For example, if the deployment was wide enough, server farms may have to be designed for load balancing as well as efficient processing of client requests.

It would also be interesting for future research to investigate the possibility of using the UpLoD architecture for the localisation of existing applications. One possibility is to develop a daemon or Windows service that would facilitate this. The daemon or the service could display translations as tooltips whenever the mouse is hovered over the UI strings of the existing application.

Future work will include the development of a suitable light-weight localisation model that includes an appropriate container that could facilitate a new and ongoing micro versioning capability. To accompany this a micro versioning workflow model would have to be developed that could facilitate and address many of the features described throughout this paper, for example the capability to facilitate a 24 hour micro update capability that could cover up to 100+ languages on a 24 hour basis.

Appropriate techniques for the development and maintenance of an associated translation history would also be a major objective. It is envisaged that these issues will be worked through, by the development of a series of prototypes for a selected sample open source application and associated user trials. This iterative design approach will then serve to inform on the overheads required to implement the UpLoD architecture and drive development of the associated tools and facilities required to optimise this approach.

References

Bookstein, A. (1990). Informetric distributions, part I: Unified overview. Journal of the American Society for Information Science 41: 368-375

Feller, J. and Fitzgerald, B. (2002) Understanding Open Source Software Development. Addison-Wesley Longman Publishing Co., Inc.

Gift, Noah and Shand, Adam (2009) Introduction to distributed version control systems, IBM DeveloperWorks, 07 Apr 2009.
https://www.ibm.com/developerworks/aix/library/au-dist_ver_control/

Howe, Jeff (2006). The Rise of Crowdsourcing. Wired. Magazine June 2006. http://www.wired.com/wired/archive/14.06/crowds.html. Retrieved 22 November 2009.

Localization World Berlin 2009.
http://www.localizationworld.com/lwber2009/about.php (last accessed 21 November 2009).

Losse, Kate (2008). Keynote at the 2008 LRC XIII Conference. Localisation4All. Dublin, Ireland, 02-03 October 2008. http://www.localisation.ie/resources/conferences/2008/keynote.htm#kate (last accessed 22 November 2009).

LRC XIV. Localisation in the Cloud. 2009 LRC Conference, Limerick, Ireland, 24-25 September 2009.http://www.localisation.ie/resources/conferences/2009/programme.htm (last accessed 21 November 2009)

Mockus, A., Fielding, R. T., and Herbsleb, J. (2000). A case study of open source software development: the Apache server. In Proceedings of the 22nd international Conference on Software Engineering (Limerick, Ireland, June 04 - 11, 2000). ICSE '00. ACM, New York, NY, 263-272. DOI=http://doi.acm.org/10.1145/337180.337209
Raymond E. S. (1999). The Cathedral & the Bazaar. O'Reilly. ISBN 1-56592-724-9. http://www.catb.org/~esr/writings/cathedral-bazaar/cathedral-bazaar/.

Raymond, E. S. (2001). The Cathedral and the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary. O'Reilly & Associates, Inc.

Rickard, Jason (2009). Translation in the Community. 2009 LRC XIV Conference. Localisation in the Cloud, Limerick, Ireland, 24-25 September 2009. http://www.localisation.ie/resources/conferences/2009/presentations/LRC_L10N_in_the_Cloud.pdf (last accessed 21 November 2009).

Schäler, Reinhard (2008). Localisation4all: Shifting the Mainstream Localization Paradigm. Localization World Conference, Berlin, 2008.
http://www.localizationworld.com/lwber2009/programDescription.php#P6 (last accessed 21 November 2009).

Yunus, Muhammad and Weber, Karl (2007). Creating a world without poverty: social business and the future of capitalism. PublicAffairs, 2007.

Guidelines for Authors

Localisation Focus
The International Journal of Localisation
Deadline for submissions for VOL 9 Issue 1 is 30 July 2010

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 and HCI, tools and technology development, cultural aspects, translation studies, human language technologies (including machine and machine assisted translation), 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.

If you wish to submit an article to Localisation Focus-The international Journal of Localisation, please adhere to these guidelines:

- Citations and references should conform to the University of Limerick guide to the Harvard Referencing Style
- Articles should have a meaningful title
- Articles should have an abstract. The abstract should be a minimum of 120 words and be autonomous and self-explanatory, not requiring reference to the paper itself
- Articles should include keywords listed after the abstract
- Articles should be written in U.K. English. If English is not your native language, it is advisable to have your text checked by a native English speaker before submitting it
- Articles should be submitted in .doc or .rtf format, .pdf format is not acceptable
- Article text requires minimal formatting as all content will be formatted later using DTP software
- Headings should be clearly indicated and numbered as follows: 1. Heading 1 text, 2. Heading 2 text etc.
- Subheadings should be numbered using the decimal system (no more than three levels) as follows:

  Heading
  1.1 Subheading (first level)
  1.1.1 Subheading (second level)
  1.1.1.1 Subheading (third level)
- Images/graphics should be submitted in separate files (at least 300dpi) and not embedded in the text document
- All images/graphics (including tables) should be annotated with a fully descriptive caption
- Captions should be numbered in the sequence they are intended to appear in the article e.g. Figure 1, Figure 2, etc. or Table 1, Table 2, etc.

More detailed guidelines are available on request by emailing LRC@ul.ie or visiting www.localisation.ie

Localisation Focus
The International Journal of Localisation
VOL. 8 Issue 1 (2009)

CONTENTS