Language Resources and Localisation

Reinhard SCHÄLER
Localisation Research Centre (LRC)
Department of Computer Science and Information Systems (CSIS)
University of Limerick
Limerick, Ireland
Reinhard.Schaler@ul.ie

Abstract

Localisation is one of the fastest growing industrial sectors in the digital world. Since the mid-eighties, the role of localisation has developed and changed dramatically. Localisation has been redefined as the *provision of services and technologies for the management of multilinguality across the global information flow*. This paper discusses the need for easily accessible dedicated language resources for localisation, provides a practical example of what can be achieved with appropriate language resources in the context of localisation and proposes a strategy to acquire, maintain and make them easily accessible.

1 Introduction

Since its emergence in the mid 1980s, localisation has largely been defined as the linguistic and cultural adaptation of products for specific locales. Over the past 20 years, Localisation has become one of the engines driving the development of the multilingual information society and probably the first industrial sector where language resources have been used widely and consistently on large-scale commercial projects.

Localisation professionals must prepare very large amounts of digital content simultaneously for different markets in acceptable quality and at affordable costs. This is only possible with the support of language resources, such as written and spoken corpora, translation memories and terminology databases, as well as the appropriate software tools for the acquisition, preparation, collection, management, customisation and use of these resources.

In the following paragraphs, we will provide an overview of the current state of the localisation industry and its requirements, and focus on those aspects of the localisation process where the successful use of language resources is crucial for the timely delivery of multilingual digital content. We will describe some of the most widely used language resources in localisation. Finally, we will describe major current research efforts relating to language resources and localisation, and highlight the opportunity for the establishment of a Language Resources Centre for Localisation.

2 The changing face of localisation

In this section, we will provide some background information which will help to explain the reasons for the changing needs of the localisation industry, particularly in relation to language resources.

2.1 Growing language populations create commercial opportunities

Any content made available on the digital networks (e.g. the internet) becomes instantly available to millions of people across the globe. To make this content accessible, however, it needs to be *localised*. In today's cyberspace, posting digital content in just one language is not sufficient anymore.

Of a total online language population of 680 million in September 2003, only 35.5% spoke

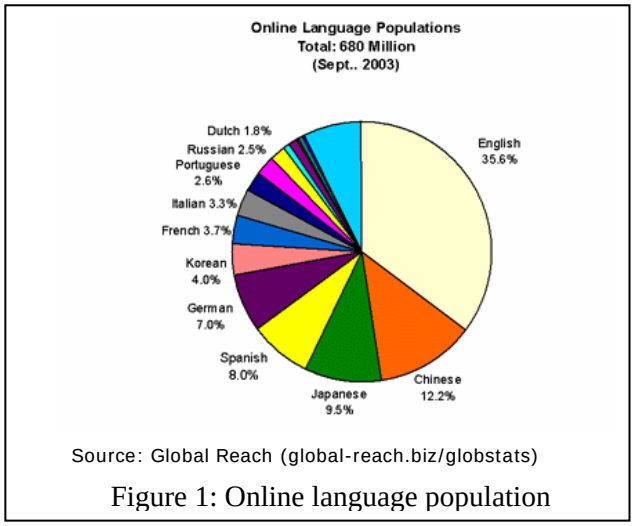


Figure 1: Online language population

English as a mother tongue, but 25% spoke Chinese, Japanese or Korean (CJK) and another

25% Spanish, German, French, Italian and Portuguese.

In this scenario, localisation is a pre-requisite for the provision of equal access to the digital information society independently of an individual's cultural and linguistic background or their geographical location and, at the same time, offers enormous business potential. In addition, it is evident that localised digital content (applications and systems) is a pre-requisite for the preservation of linguistic and cultural diversity in the digital world.

According to US-analysts Allied Business Intelligence, the world-wide market for translation and software or web localisation is growing from US$11 billion (1999) to US$20 billion this year.

However, although much digital content is created in Asia and Europe, 95% of localised digital content still originates in the USA.

2.2 Localisation services – redefined

While politicians all over the world want to make Information Society Technologies (IST) available and accessible in the language and locale of the people they represent, software and digital content publishers need to respond to the demands of their customers by supporting a wide variety of local languages and cultures in their products.

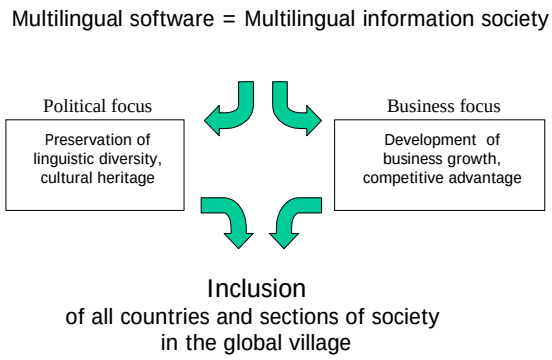


Figure 2: Multilingual software and Infosociety

Since its emergence in the mid-1980s, the localisation industry has taken on the task of responding to these requirements of business and politics. Initially seen as just one of many service suppliers to the general IT sector, it is now taking on a more independent role translating the global digital content challenge into new business and social opportunities. It is the localisation industry that can enable the open, pluralist, user-friendly and inclusive multilingual and cross-cultural information society.

The dramatic change in the role and the funtion of localisation has happened in parallel with the development and changes of the IT and content publishing sectors in general, now all converging in the digital world.

The businesses of computing (hardware, software and services), communications (telephony, cable and satellite), and content (publishing, entertainment, advertising) are coming together to create the *new digital media industry*. New media publishing on the Internet combines computing, entertainment, broadcasting, music and

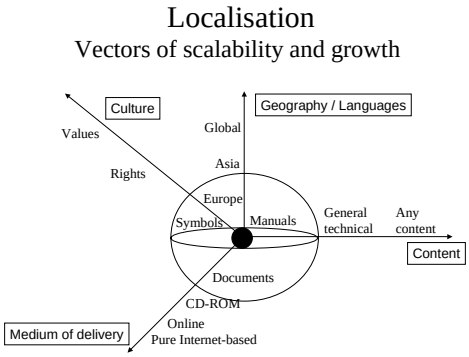


Figure 3: Vectors of scalability and growth

video production.

The issues faced by a wide variety of formerly independent, unconnected traditional content publishing industries joining in the digital world include the need to handle, control and translate larger amounts of text than ever before into an ever increasing variety of languages in parallel with the development of the original version, within a tight budget and according to strict quality guidelines as well as the need to adapt — not just translate — their products to the culture and locale of the target market.

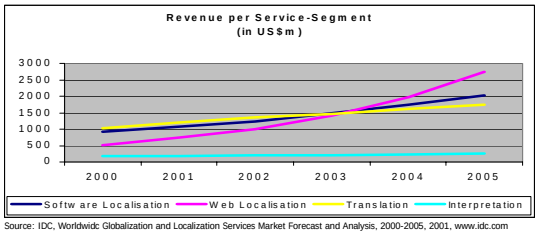
IT provides the framework for the convergence of these activities. The localisation industry provides the framework for the convergence of the multilingual aspects of these activities. Localisation becomes the catalyst for electronic multilingual production and publishing.

On the background of these developments, the concept of localisation is being redefined as the *provision of services and technologies for the management of multilinguality across the global information flow*.

Timely and cost effective delivery of high quality digital content to the global marketplace has become the major growth area for the localisation industry. It opened the relatively narrow software localisation industry to a wider range of players who are broadening traditional roles within the software localisation industry.

Yet, the localisation industry does not have access to a robust infrastructure comprising

- Start: mid-eighties
  - Packaged softward -> multimedia -> content
- Ireland: the world centre (certainly the European centre)
- 95% of source orginates in the USA
- International market more important for pubishers than domestic markets
- MS: >60%, >US$5b ., >1,000 projects/year, Ireland: US$1.9b revenue (2001)



Source: IDC, Worldwide Globalization and Localization Services Market Forecast and Analysis, 2000-2005, 2001, www.idc.com

Figure 4: Revenue per service segment

language data and tools, which are a prerequisite for the timely and cost effective creation and deployment of multilingual, cross-cultural and multimodal digital content. There is an urgent need for the development of the structural basis to make a sustained internationalisation and localisation effort possible, especially for less widely spoken languages.

2.3 Localisation research

The new role of the localisation industry also creates new opportunities and requirements for research and development.

How can the business requirement for a reduction in production cost, combined with fast throughput time and high quality be satisfied in the context of localisation?

One answer to this question is by access to adequate language resources, tools and standards.

Re-use of already translated material, translation *recycling*, can not only speed-up translation, it can also at least help to ensure a more consistent use of terminology and thus increase the quality of the translation.

Access to adequate terminology tools and resources can also help to achieve better translations at a lower cost.

To make appropriate use of these technologies and resources, their role must be assessed in the context of localisation.

3 Language resources and localisation

Language data or resources have been defined as a set of speech or language data and descriptions in machine readable form, used e.g. for building, improving or evaluating natural language and speech algorithms or systems, or, as core resources for the software localisation and language services industries, for language studies, electronic publishing, international transactions, subject-area specialists and end users.

Examples of language resources are written and spoken corpora, computational lexicons, terminology databases, speech collection and processing, etc. Basic software tools are also important for the acquisition, preparation, collection, management, customisation and use of these language resources and other resources. (see: http://www.elra.info/article.php3?id_article=35, consulted May 2004)

Currently, there are no large collections of language data relevant to localisation easily accessible. While individual digital content developers, especially large multinationals, have very large collections of multilingual language data available and use them in a highly effective and efficient way, these data collections are not available to the localisation community in general.

Large multinational content publishers, however, have shown how the efficient use of language resources can help to achieve astonishing production results leading to what some have described as the *translation factory*.

3.1.1 Case study

Tony Jewtushenko, Tools Manager with Oracle's Worldwide Translation Group, presented the following example of the use of language resources in localisation at the LRC's 2003 Annual Localisation Conference.

Project constraints

- 4m wordcount software strings;
- 30 languages simultaneous release;
- 13k localisable files;
- Localisation group in Dublin; 5,000 people world-wide distributed development team.

Objectives

- 24/7, 100% automated process – no exceptions
- Translation in parallel with development
- Translation begins at code check-in
- Translation "on demand" – no more "big project" model

Solution: the translation factory

- Current throughput: 100,000 language check-ins per month
- 2 million files per month
- 98% of words leverage
- Average time to process a file: 45 seconds
- Fully scalable "add-a-box model"
- Simpship of all 30 languages
- International version testing before US release
- Reduced no. of release engineers (20->2) resulting in US$20m saving per year
- Positive ROI within 1 year

It is important to keep in mind that Oracle is one of the world's leading digital publishers and runs one of the most sophisticated internationalisation and localisation operations in the world. Oracle is also centrally involved in the development of two

key standards under the umbrella of OASIS, the XML-based Localisation Exchange Format (XLIFF) and the Translation Web Services Group (TWS), which, combined, have the potential to fundamentally change not just the way localisation is done by Oracle, but by every digital publisher bringing its contents to the global market.

3.1.2 Support for SMEs and researchers

There are no good reasons why the work of other organisations, such as small and medium sized enterprises (SME's) and research organisations, should not benefit from the intelligent use of linguistic resources, including language data, tools and standards. While large organisation catering for the main language markets have access to the finances necessary for the development and maintenance of this linguistic infrastructure, smaller organisations and those catering for financially less significant markets will need the support of a shared and widely supported infrastructure.

This infrastructure would need to cover:

- Multimodal digital content in source and target languages;
- Monolingual and multilingual terminology;
- Translation memories.

3.2 Linguistic tools

Linguistic tools are seen by the Localisation Industry Standards Assocication (LISA) generally as still an emerging sector (*Localization Industry Primer)* although, according to LISA, enormous progress has been made over the past years in the area and a number of productivity enhancing tools are now in use, without which the localisation industry as we know it today would not be able to operate.

3.2.1 Current situation

The issues, which are addressed by linguistic tools and technologies answer some of the central questions asked by localisation professionals around terminology handling and update processing.

Terminology handling

- Where can translators find standard terminology in multiple languages?
- How can multilingual terminology be processed so that it can be made readily available and easily accessible?
- Are there feasible models and mechanisms to maintain and constantly update multilingual terminologys so that modifications can be made accessible to translators instantaneously?
- How can changes in previously agreed terminology be automatically integrated in already translated text?
- How can translated texts automatically be checked for the correct and consistent use of terminology?

Update processing

- How can versions of the same source material be compared against each other automatically?
- How can overlaps be identified, marked and analysed?
- How can source and target language content be compared and aligned?
- How can already translated text fragments (exact or fuzzy matches) automatically be used for the generation of a new target version?
- How can the limitations of *Translation Memory Systems* be overcome?

The linguistic tools and technologies most widely needed and developed for use in the internationalisation and localisation effort include:

- Terminology management systems, which aid the collection and use of specialised vocabularies;
- Translation memories, which are designed to facilitate the reuse of previous translations;
- Machine translation, which provides actual linguistic analysis and conversion of texts from source language into the desired target language;
- User interface and user assistance visual translation environments, which aid translators to interactively work with compiled and uncompiled resource files in a variety of formats;
- Language data analysis tools, which rapidly compare and analyse old and new source material;
- Sophisticated matching tools leveraging material from previous projects;
- Natural language parsers;
- Extract-and-Insert tools;
- Parsers for natural language digital content in compiled sources.

While large and sophisticated localisation operations have easy access to relevant linguistic third party technologies and in-house tools, smaller operations often do not. Reviewing the large variety of sophisticated tools and technologies available on the market, they often shy away from the purchase and implementation of tools because of the perceived high-risk factor attached to their deployment.

3.2.2 Impact

A support infrastructure must be put into place to allow smaller players involved in localisation direct and online access to the widest variety of tools and technologies and detailed information about these. A first step in the implementation of this infrastructure has been the establishment of the Localisation Tools and Technology Laboratory and Showcased (LOTS) as part of the European-funded ELECT project.

Cooperating with leading industry associations, such as the Globalisation and Localisation Association (GALA) and The Institute of Localisation Professionals (TILP), and building on the expertise available within the ELECT consortium, a sophisticated online library with detailed background information on each of the tools available was prepared and published.

The detailed LOTS-sponsored reviews of individual tools and technologies will allow potential users of linguistic tools and technologies, specifically those working in small and medium sized enterprises, to base their decision on which tool to use for their particular localisation needs on independent, well-researched and easily accessible information.

3.3 Standards

Standards played a central role in the establishment of the localisation industry's first association, the Localisation Industry Standards Association (LISA). Very early on, localisation professionals recognised that the successful implementation of widely recognised standards would lower the cost of localisation, shorten the time necessary for the successful completion of projects and increase the quality of the products delivered to international audiences.

The following graphic visualises the role standards for linguistic resources play in the overall localisation process (source: i18n Inc.: www.i18n.ca):

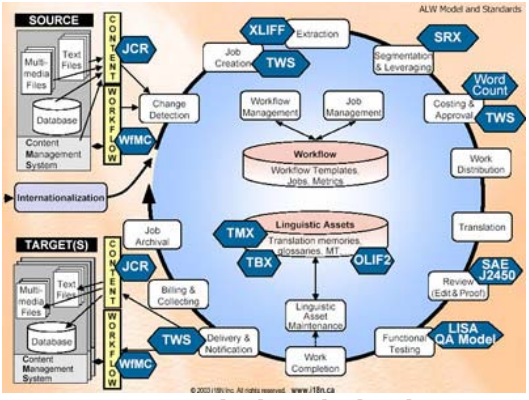


Figure 5: Standards in the localisation

3.3.1 Current situation

A large number of standards relevant to linguistic resources in the context of localisation have been published by a number of organisations identified by LISA as being involved in the development of standards. Among these are:

International Standards Organisation (ISO) – This is a network of national standards institutes from 140 countries working in partnership with international organizations, governments, industry, business and consumer representatives. The ISO sees itself as a bridge between public and private sectors.

- Technical Committee 37 - Terminology and other language resources
- ISO 639 - Language Codes
- Terminology Data Categories - ISO 12620
- MARTIF - ISO 12200 - Machine-readable terminology interchange format
- Terminology Work - ISO 704
- Vocabulary - ISO 1087-1 - Part 1: Theory and Application
- Vocabulary - ISO 1087-2 - Part 2: Computer Applications
- Terminological Markup Framework - ISO DIS 16642
- ISO639-1 : New ISO standard for the identification of languages names

Localisation Industry Standards Association (LISA) – This organisation has published a variety of standards relevant to the use of language resources in localistion:

- TMX the exchange standard for translation memory data between tools and/or translation vendors aiming at little or no loss of critical data during the process.
- TBX the open XML-based standard format for terminological data.
- OLIF the XML-compliant standard for terminology offering support for natural language processing (NLP) systems, such as machine translation, by providing coverage of a wide and detailed range of linguistic features.

OASIS – This international, not-for-profit consortium designs and develops industry standard specifications for interoperability based on XML. Two of these have been developed specifically with localisation in mind:

- XLIFF – the XML-based Localisation Interchange File Format
- TWS – the Translation Vendor Web Services Standard

The Free Standards Group Open Internationalization Initiative (Openi18n.org) – This non-profit initiative aims to accelerate the use

and acceptance of open source technologies through the application, development and promotion of interoperability standards.

Termnet – The International Network for Terminology promotes co-operation in the field of terminology internationally, so as to stimulate the development of the terminology and knowledge market, as well as terminology proper. Termnet publishes terminologically relevant data in both printed and computerised forms and thus makes it accessible to a large circle of users.

Unicode – The Unicode Consortium is a non-profit organization founded to develop, extend and promote the use of the Unicode Standard, which specifies the representation of text in modern software products and standards. The membership of the consortium represents a broad spectrum of corporations and organisations in the computer and information processing industry. Membership in the Unicode Consortium is open to organisations and individuals anywhere in the world who support the Unicode Standard and wish to assist in its extension and implementation. Unicode's most visible activities include the holding of the Internationalisation and Unicode conference twice a year.

WC3 – This consortium develops interoperable technologies (specifications, guidelines, software, and tools) to lead the Web to its full potential as a forum for information, commerce, communication, and collective understanding.

3.3.2 Impact

Currently, there is not central repository of standards relevant to the development and maintenance of linguistic resources for localisation comprising language data, tools and standards which is easily accessible to the localisation community. Furthermore, and equally important, no independent organisation or consortium is currently set up to demonstrate the effective and efficient use of linguistic resources in a localisation environment following industry-standard approaches and using state-of-the-art technologies.

4 The Localisation Tools, Technologies and Resources Laboratory

4.1 The Rationale

The establishment of the Localisation Tools and Technologies Laboratory and Showcase (LOTS) as part of the European-funded ELECT project was the first attempt to make a repository of language resources covering linguistic data, tools and standards available and easily accessible.

Although LOTS is located at the Localisation Research Centre in Limerick, it is also available online via www.electonline.org.

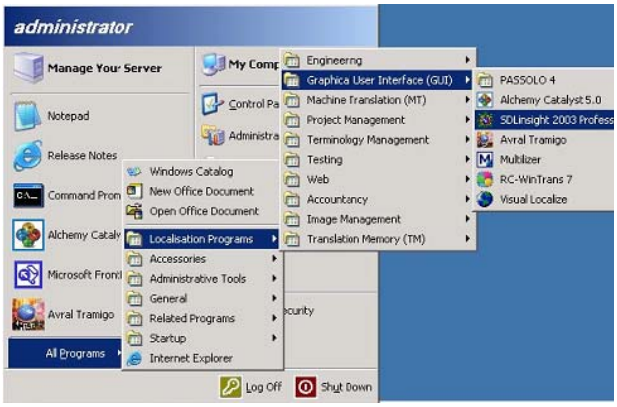


Figure 6: The LOTS online desktop

In one location, LOTS provides different user groups with access to the widest possible range of tools and corresponding resources.

- Students, Trainers can quickly get an overview of tools, technologies and resources relevant to localisation.
- Researchers can experiment with state-of-the-art technologies and resources comparing the results of their efforts with commercial offerings.
- Professionals can test whether particular applications are appropriate to cover their specific needs.
- The LRC uses the facilities available in LOTS to verify standards and interoperability issues.

4.2 Tools and Technologies

LOTS was established with the support of the localisation tools and technology developers. All the resources available to LOTS have been given to the LRC by their owners free-of-charge.

Twenty-four companies, resprenting the majority of localisation tools and technologies worldwide, have so far contributed to LOTS.

Figure 7: The LOTS laboratory

The LRC has estimated that the laboratory contains tools and technologies worth more than €350,000.

These are accessible on a large number of PCs running in LOTS under a variety of operating systems in more than a dozen different languages.

They are also available online on the LOTS server which brings LOTS directly to the desktop of users anywhere in the world.

4.3 Resources

In addition to the tools and technologies, the resources available on the LOTS server and in the laboratory include the most widely used file formats and standards.

While tools and technologies are well covered in LOTS, the coverage of corresponding resources as used by these tools could be improved.

The LRC is working with content and technology developers on agreements which will allow a wider deployment of authentic source material in a wider range of languages and file formats.

5 Language Resource Centre for Localisation

Although with LOTS the LRC has established the foundations for a language resources centre for localisation, a consortium representing the main actors in localisation, based in a number of European countries, including the accession countries, will need to be established to develop and maintain such a repository long-term. This consortium will also need to develop long-lasting relations with the leading industry associations and standards bodies, e.g. the Globalisation and Localisation Association, LISA, OASIS, Unicode, W3C and The Institute of Localisation Professionals (TILP).

The centre, proposed to be located at the LRC, will provide access to

- Linguistic resources
- Methodologies and guidelines
- Localisation scenarios

Each of these components will be explained more in detail in the following sections.

5.1 Repository of linguistic resource

The centre will establish, develop and maintain a repository of linguistic resources for localisation physically based at the LRC but accessible online over the Internet covering:

- A large variety of multimodal authentic digital content in source and target languages;
- Mono- and multilingual terminology;
- Translation memories;
- Linguistic tools and technologies used for the automatic processing of digital content;
- Guidelines and standards for the development and processing of digital content to be localised.

Together, the efforts in these areas will deliver the structural basis for a sustained internationalisation and localisation effort, especially for less widely spoken languages where market forces often provide insufficient incentives.

5.2 Methodologies and guidelines

The centre will provde access to methodologies and guidelines for the verification of standards compliance and interoperability verification for linguistic resources in a multilingual, multicultural and multimodal localisation environment. This work will be based on industry-standard approaches and be guided by established principles and procedures. The reports will be sourced in cooperation with relevant associations and standards bodies and public deliverables of European Union funded projects.

5.3 Localisation scenarios

The centre will, in consultation with the wider localisation community, build a laboratory-based, automated localisation environment mirroring real-world, authentic localisation scenarios. This environment will be used to showcase, verify and demonstrate best practice in localisation making use of the linguistic resources available through the centre.

Results from each of these three areas of activity will be made available to the wider digital content and localisation communities using different dissemination strategies. It is envisaged that the development of a market place for these linguistic resources will guarantee the sustainability of the effort.

6 Conclusion

We have shown how the needs and requirements of localisation have developed over the years. The enormous pressure on localisation providers to produces language versions of original material simultaneously with the production and publication of the original can only be addressed making efficient and effective use of customised language resources covering linguistic data, tools and technologies, and appropriate standards.

Access to language resources for SMEs and the research community can be realised within a widely supported Language Resource Centre for Localisation, built on the foundations of the Localisation Tools and Technology Laboratory (LOTS) at the LRC.

The establishment of the Language Resources Centre for Localisation is a five-year project. During this time, the LRC plans to create a sustainable, accessible and financially viable linguistic infrastructure for the internationalisation and localisation communities.

Its overall aims are to:

- Pool together linguistic infrastructure resources for the localisation industry, including digital content, monolingual and multilingual terminology, translation memories, tools and technologies, as well as relevant standards.
- Establish a linguistic resources support network within the localisation industry covering digital content publishers, service providers, technology developers, standards bodies and standards verification initiatives.
- Provide convenient access to relevant linguistic resources for content developers, service providers, as well as suppliers of localisation services and solutions.
- Develop methodologies for the verification of standards compliance.
- Implement localisation scenarios in a laboratory environment aimed at demonstrating state-of-the-art, best practice localisation technology and process solutions, and at verifying relevant localisation standards.
- Work towards the establishment of a market place and a viable linguistic resources provider network for the localisation industry.

Activities during years 1-2 will focus on the establishment of a repository of linguistic resources. During this period, the LRC will build a core group of partners supporting its development. This core group will be backed by the LRC's Contact Group.

Activities during years 3-5 will see the implementation of a financially viable standards testing and interoperability verification system at the LRC. While few commercial organisations have managed to make a profit to sustain this kind of activity in other industrial sectors, there is no doubt that the operation of such a centre is financially viable long-term in a not-for-profit environment such as that provided by the LRC at the University of Limerick.

7 Acknowledgements

We would like to acknowledge the support of the European Union for the European Localisation Exchange Centre (ELECT) under its eContent Programme, contract 52005.

References

Electonline – the online resource for the localisation community. www.electonline.org, last visited 10 June 2004.

Language resources: the ELRA definition. http://www.elra.info/article.php3?id_article=35, last visited 20 May 2004.

Localization Industry Primer. LISA. www.hltcentral.org/usr_docs/call_docs/ eContent/call1/LISA%20Primer.pdf, last visited 13 May 2004.

Schäler, R., Michael Carl, Andy Way. 2002. *Toward a Hybrid Integrated Translation Environment.* In: "Proceedings of the Fifth Biennial Conference of the Association for Machine Translation in the Americas Conference (AMTA)", Tiburon, California, 08-12 October 2002.

Schäler, R. 2002. *The European Localisation Exchange Centre* (Keynote). In: Proceedings of the Twenty-first International Unicode Conference, Unicode Localization and the Web: The Global Connection, Dublin, Ireland, 16-17 May 2002.

Schäler, R. 2002. *The XLIFF Standard.* Panel Session with Ian Dunlop (Novell), Tony Jewtushenko (Oracle), Wojtek Kosinski and Peter Reynolds (Bowne), in: Proceedings of the Twenty-first International Unicode Conference, Unicode Localization and the Web: The Global Connection, Dublin, Ireland, 16-17 May 2002.

Standards in the localisation process. Graphical representation. www.i18n.ca, last visited 10 May 2004.