	<p>Intera</p> <p>REPORT ON THE MODEL OF LRS PRODUCTION</p>
---	--

<i>Project reference number</i>	e-content EDC-22076 INTERA / 27924
<i>Project acronym</i>	INTERA
<i>Project full title</i>	Integrated European language data Repository Area
<i>Project contact points</i>	Khalid Choukri Evaluations & Language Resources Distribution Agency S.A. 55-57 Rue Brillat Savarin, 75013 Paris, France Phone: +33 1 43 13 33 33, Fax: +33 1 43 13 33 30 Email: choukri@elda.fr
<i>Project web site</i>	http://www.elda.fr/intera
<i>EC project officer</i>	Philippe Gelin
<i>Document title</i>	Report on the model of LRs production
<i>Deliverable ID</i>	D 5.3
<i>Document type</i>	Report
<i>Dissemination level</i>	PP
<i>Contractual date of delivery</i>	
<i>Actual date of delivery</i>	
<i>Status & version</i>	Final, v. 1.1
<i>Work package & task ID</i>	WP5
<i>Work package, task & deliverable responsible</i>	ILSP
<i>Number of pages</i>	22
<i>Author(s) & affiliation(s)</i>	Maria Gavrilidou, Voula Giouli, Elina Desipri (ILSP) Monica Monachini, Claudia Soria (ILC)
<i>Additional contributor(s)</i>	
<i>Keywords</i>	LRs production methodology, business model
<i>Abstract</i>	The present deliverable presents a business model for the production of language resources, based on the experience the project's partners had within WP5. This model aims to suggest methodologies and stages of the LRs production, that would make the process attractive to LRs producers.
<i>Additional notes & remarks</i>	

Table of contents

2EXECUTIVE SUMMARY.....	3
3INTRODUCTION.....	3
4THE INTERA EXPERIENCE	4
4.1TARGET GROUP IDENTIFICATION.....	4
4.2USER NEEDS AND REQUIREMENTS.....	4
1.1.1Languages.....	4
1.1.2Domains.....	5
1.1.3Sources targeted.....	6
1.1.4Processing specifications.....	6
4.3PROBLEMS FACED.....	6
1.1.5Availability problems	7
1.1.6The Web as a source.....	7
1.1.7Technical problems	9
5MARKET NEEDS.....	10
5.1BUSINESS PRACTICES RELATED TO LRS.....	11
1.1.8Are LRs needed?.....	11
1.1.9How are LRs acquired?.....	11
1.1.10Which languages are mostly needed?.....	11
1.1.11Raw or processed data?.....	12
1.1.12Metadata descriptions.....	13
1.1.13Are standards crucial?.....	13
1.1.14Texts without processing tools?.....	13
1.1.15Other issues.....	14
6BUSINESS MODEL FOR THE LR PRODUCTION COMMUNITY.....	14
6.2BUSINESS MODEL: DEFINITIONS.....	14
6.3LR PRODUCTION: THE BUSINESS	15
6.4THE BUSINESS MODEL	16
1.1.16Market-oriented development of LRs.....	17
1.1.17LR production methodology	18
1.1.18Text processing tools.....	19
1.1.19Resource configuration.....	19
1.1.20LR portability and re-usability.....	20
1.1.21Dissemination and promotion strategies that could apply to LR producers	20
7CONCLUSIONS.....	21
8REFERENCES.....	22

2 Executive Summary

This deliverable presents a model for the LRs production based on the experience of the INTERA production task. It discusses the procedure followed within the project, focuses on the problems faced which had an impact on the initial goals, presents the necessary modifications that resulted from these problems, evaluates the market needs as attested by various surveys, describes the methodology that is proposed for the efficient Language Resources Production and finally concludes with suggestions that could make the LRs production process a successful endeavour attractive to LRs producers.

3 Introduction

The aim of WP5 of the INTERA project was not simply to produce new LRs but also to suggest ways and techniques in the LRs production business that can be exploited in order to make it a profitable task attractive to the eContent professionals and – ultimately - to other possible users such as IT developers, etc. The idea is to show that the endeavour of LR production can be undertaken in such a way that it is proportionate to the users' needs and, thus, profitable. To this end, WP5 makes a first step in the construction of the required infrastructure, whereas, on the other hand, it prepares the grounds for its further development, by outlining the processes and principles according to which such endeavours should be undertaken.

Therefore, the perspective adopted for the proposed model for LR producers takes into consideration (a) the actual status of LRs in terms of sheer existence, availability, language coverage, format, annotation as well as (b) the actual market needs for LRs and for relevant processing tools. This particular point of view inevitably brings us to try to conciliate two different, opposite forces. On the one hand, there is the need to establish a reference model for multilingual resources which is up-to-date and compliant with existing standards. From this point of view, any prospective resource should aim at completely satisfying user needs and requirements, as well as complying with existing standards. On the other hand, we must consider the actual viability and feasibility of language resources production, especially for those languages which suffer from poor representation and for which raw material is scarce. This means that a realistic production model should also take into consideration very basic problems of data availability, representativeness, and size, together with availability of language processing tools.

In previous deliverables (D5.1) we discussed the results of the survey on the eContent specialists' needs in terms of the *domains* and the *languages* that are of most interest to eContent professionals and that we should, therefore, focus on. In deliverable D5.3 we will make an attempt to further analyse the findings of the survey in order to better define the eContent market requirements. We will then reconsider the users' needs attested in combination with the existing practices and policies regarding the LRs production and the methodology we followed in the framework of the WP5. Finally, we will conclude to a proposed model on the basis of information gathered from all the aforementioned sources in line with user needs and eContent market requirements.

4 The INTERA experience

The following section briefly describes the LR production process as implemented within the INTERA project. This will serve as a basis for the proposal for a business model.

The approach adopted in building the INTERA parallel multilingual corpus has been an hybrid one. Initially, a **user-oriented** approach has led to the identification of multilingual LR specifications on the basis of (a) the current situation in the LR area, and (b) user needs and requirements on the basis of the professionals' working habits and processes. However, shortcomings in the acquisition and processing of the data, has led us to the adoption of a more **realistic scenario**, with the back-off solution prevailing, i.e., the creation of pairs of parallel corpora in the decided languages with English serving as the pivot. This methodology is described in the following sections.

4.1 Target group identification

More specifically, early in the project life cycle, INTERA was focused on eContent professionals and the selection of the type of resources was performed on the basis of their identified needs and requirements. The target group of eContent players addressed by the project (see also Deliverable D5.1) has been further defined as consisting of professionals involved with the:

- production of digital content (authors or publishers)
- Globalisation, Internationalisation, Localization and Translation (GILT) processes, and
- development of Human Language Technology (HLT) software, ranging from multilingual information retrieval and extraction tools, to content management and Computer-Assisted Translation or Machine Translation solutions.

4.2 User needs and requirements

The next step concerned the identification of user needs and requirements on the basis of the professionals' working habits and processes. This was achieved by exploiting the results of a number of previous initiatives to roadmap the state-of-the-art in multilingual LRs, in combination with new initiatives undertaken in the framework of the project and targeted to the eContent world (see also D5.1). The information elicited from the surveys mentioned above was coupled by a study of the activities of the eContent professionals as regards LRs, conducted in the framework of INTERA (Gavriliidou et al, 2004) through the circulation of a questionnaire distributed to potential users, as well as through personal contacts with a number of actors in the relevant fields. The main areas of the study concerned the types of LRs the eContent professionals are interested in, domains and languages of interest, and, most important, policies concerning the way they acquire, use and exploit LRs and tools.

The study on the needs of the target group yielded the specifications according to which the LRs production process was implemented. These are the objective of deliverable D5.1, where they are extensively described – here they are briefly discussed.

1.1.1 Languages

The project aimed at the development of LRs for “less widely spoken languages”. Given that the INTERA project belongs to the eContent programme, the term “less widely spoken” has been interpreted as “less widely available in the digital world”; as being more appropriate to

the programme's goals and objectives. As attested, English is indeed the most used language, yet a growing tendency to use other languages is evidenced. These include not only the other competing "international" ones but, most important, even the less widely spoken national languages. It is becoming increasingly evident that people prefer to communicate in their own languages. This demand puts further burdens on the eContent world, since products and services must be presented in a growing number of languages in order to be marketable. A robust infrastructure comprising language data and tools is a prerequisite for the quick and cost effective creation and deployment of multilingual and cross-cultural eContent. It is the structural basis for any sustainable internationalisation and localization effort, especially for the less widely available languages where market forces often provide insufficient incentives. In order to promote the multilingual and multicultural eContent, the LR community should focus on languages having still few LRs describing them, so as to facilitate the dissemination of eContent in these languages by the eContent experts.

Taking this point into account, the INTERA project has focused on the creation of multilingual resources for the less widely available languages. This was further reinforced by the user needs study, which pointed towards the less widely represented languages (exactly because of the scarcity and the difficulty of locating high-quality language material). Given the accession of the new countries into the EU and the prospective accession of more countries of the Central and Eastern European area, the focus of the INTERA effort was on the Balkan languages, and specifically on Bulgarian, Serbian and Slovene; Greek was also included due to its under-representation in the digital market and due to the difficulties stemming from the linguistic idiosyncrasies (non-Latin character set, separate language family etc.).

The ideal scenario for the intended application of term extraction would be that of having a corpus with a source or pivot language and translations of the same texts in a number of target languages; however, given that the project aimed at proposing realistic solutions to be adopted in the future by prospective LRs creators, real-life drawbacks should be taken into account. Consequently, in spite of the initial target to produce a multilingual parallel corpus by collecting texts that appear in all the languages involved, scarcity of really multilingual textual data and availability limitations of the relevant processing tools imposed **the modification of the initial goal**. In this context, we opted for the creation of **comparable corpora**, in the sense of producing sets of **bilingual parallel corpora in specific domains**.

1.1.2 Domains

Similarly, the selection of the INTERA domains was also driven by user needs and requirements. The total corpus is divided into sub-corpora, each one covering one domain of the ones identified as important to the eContent professionals by means of a user survey conducted by the project participants (see INTERA Deliverable D5.1). The responses to the questionnaire coupled with information available over the Internet showed that eContent specialists move across a variety of domains depending on the application; obviously, eContent professionals are more interested in specialized domains than in general language resources

The domains prevailing are the following: health and medicine, tourism, education, law, IT/telecommunications, automotive industry etc. In the framework of the project we decided to focus on the prevailing domains as long as they promote multilingual and multicultural content: health, tourism, education and law. These further correspond to the predominant digital activities, namely, eHealth, eTourism, eLearning, eGovernment and eCommerce.

1.1.3 Sources targeted

The textual material collected mainly originates from the Internet, provided by organizations/institutions that wish to make their own material available in more than one language. Among the various sources for obtaining multilingual data, we opted for national and international organizations (e.g. United Nations, European Union, World Health Organization, etc.), Non-Governmental Organizations, multinational companies, companies with activities outside their own country (e.g. data describing company profiles & activities, product catalogues, etc.), public administration services (e.g. regarding bilateral agreements, regulations for immigrants, etc.), news agencies (targeting international broadcasting or for foreign language audience within their own country), official national government sites, national tourism organizations, translation houses, etc. In all the above cases, the material consists of either web content per se (i.e. mainly bilingual web sites, rarely trilingual or quadrilingual) or of texts (official documents, technical reports, etc.) included in the web sites.

1.1.4 Processing specifications

The specifications for the processing of the corpus were decided on the basis of the requirements of its intended application, which is the *extraction of terminology*. These specifications involve the following tasks:

- ★ *alignment* of the texts
- ★ external and internal *structural annotation* (segmentation at sentence level for the alignment task and metadata information that will be required for the description of the corpus in order to facilitate future distribution and re-use of the corpus;
- ★ *linguistic processing* (below-Part of Speech (PoS) tagging and lemmatisation)

To ensure re-usability of the collected and processed material, compliance with the following internationally accredited standards was also decided:

- ★ the aligned material conforms to the *TMX standard* (Translation Memory eXchange, <http://www.lisa.org/tmx/>), which is XML-compliant.
- ★ for the external annotation, the *IMDI metadata schema* (IMDI, Metadata Elements for Session Descriptions, Version 3.0.4, Sept. 2003, http://www.mpi.nl/world/ISLE/schemas/schemas_frame.html) has been selected;
- ★ the internal structural annotation adheres to the *XCES standard*, i.e. the XML version of the Corpus Encoding Standard (XCES, <http://www.cs.vassar.edu/XCES/> and CES, <http://www.cs.vassar.edu/CES/CES1-0.html>).
- ★ the linguistic annotation of the texts also adheres to the *XCES standard*, which incorporates the EAGLES guidelines for morphosyntactic annotation (<http://www.ilc.cnr.it/EAGLES96/home.html>).

4.3 Problems faced

The following section describes the problems faced during all phases of corpus production. The aim of this problem reporting is to actually describe the real-life situation in what concerns resource production, which will lead to the proposal for a business model (objective of deliverable D 5.3) reflecting as accurately as possible the obstacles faced in such an endeavour.

1.1.5 Availability problems

One of the major problems INTERA participants had to cope with was that of the availability of multilingual material (resource and relevant processing tools). The survey conducted in the framework of the INTERA project (D 5.1) confirmed the presumption that few resources are available either multilingual or even monolingual for the less widely represented languages.

The prominent obstacle of availability of multilingual resources and of relevant tools is strongly related, on the one hand, to the languages targeted by the INTERA project, which suffer from poor representation and for which raw material is scarce. On the other hand, policies adopted as far as the production of LRs is concerned, do not really assist the representation of the specific languages in the digital world.

1.1.6 The Web as a source

Internet has been the main source for obtaining the textual data. However, one of the most significant problems concerning the identification of multilingual LRs conformant to the specifications of the INTERA project as described in the deliverable D5.1, is the representativeness of the targeted languages over the digital content, namely over the Internet. According to figure 1 (source: <http://www.gtreach.com/globstats/index.php3>) the English language still is predominant as regards the online language populations. Yet only 5.4% - little more than 1 in 20 - of the world's population have English as their mother tongue, according to IDC (the Boston-based IT market researchers).

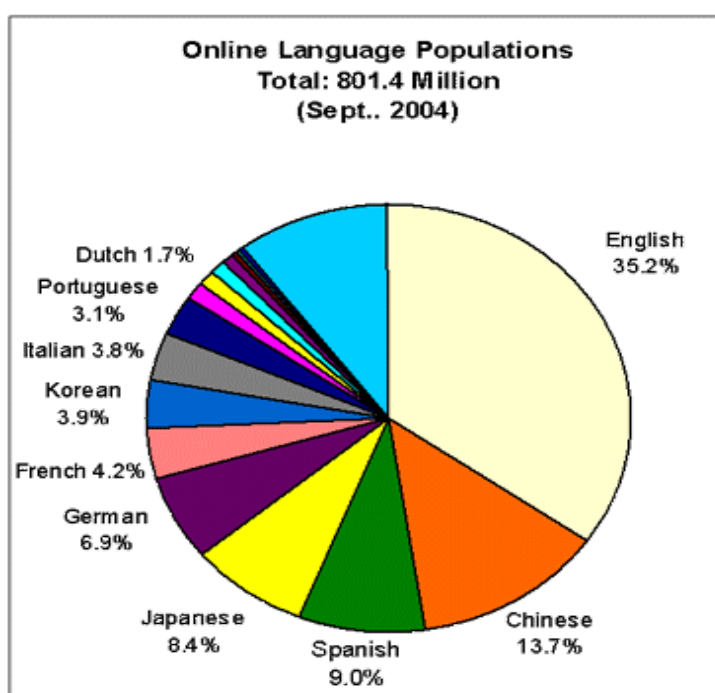


Figure 1. Online language populations

The identification of multilingual LRs in the framework of WP5 of the INTERA project revealed a sort of “paradox” as regards the use of languages and their representativeness in the digital content. More specifically, although we tend to talk about “less widely spoken languages” when referring to eastern European languages and Balkan ones, a significant

number of these languages appear among the forty most widely spoken languages all over the world (source: <http://www.globallanguages.com>). However, these languages seem to be rather underrepresented as regards the digital content. These remarks have led to the re-definition of "less widely spoken languages" to that of "less widely available languages". These languages are also known as "low density languages" due to the lack of digital resources.

The identification of existing material in the languages of interest has also unveiled the real status of the web, which is attested to be "strangely" multilingual. The representation of the targeted languages in the so-called multilingual sites has proven to be non systematic at all. Even if international organizations and multilingual portals claim to provide their content in a variety of languages, they usually provide it in the most dominant ones, namely English, French and Spanish, disregarding the less widely available languages, targeted by the INTERA project. As also stated in Resnik and Smith (2003), parallel corpora "unfortunately [...] are not readily available in the necessary quantities."

Apart from the policy of the international organizations concerning the presence of the various languages in their multilingual digital content, the identification of multilingual LRs was also impeded by the policy of Balkan organizations as regards the production of multilingual LRs. More specifically, as also attested by previous surveys (i.e. ENABLER, Gavrilidou, M., E. Desipri (2003), Final Version of the Survey, ENABLER Deliverable 2.1), Balkan organizations provide LRs for their own languages and, when creating multilingual resources, they combine them mainly with the major European ones. This situation forced us to early abandon the ideal scenario of the production of true parallel corpora, namely one (or more) text(s) translated in all languages. The modification of the initial goal was considered inevitable and consequently we were re-directed to the scenario of sets of comparable corpora in the sense of sets of bilingual parallel corpora in the specified domains.

In certain cases, the non-availability of texts was also due to legal problems such as IPR and copyright issues; in these cases, the acquisition of the texts demanded clearance of these issues. In case of failure in the clearance process, the texts had simply to be abandoned.

As regards terminological resources, the non-availability of texts entailed the elaboration of different "back-up solutions" for their production.

As already introduced in Deliverable 5.1 about Technical Specifications and refined in Deliverable 5.2 about Resource Production, different scenarios can be envisaged, especially when working with under-represented languages:

Scenario no. 1 amounts to having a true parallel corpus, i.e. one or more texts translated in all languages. Moreover, the corpus is annotated for linguistic information such as grammatical categories of words and reference form (or lemma). This scenario represents an "ideal" situation, where the extraction task can be performed working on one language which acts as a pivot language. This pivot language preferably is a language for which a lot of NLP tools and lexical resources are available. Moreover, since the corpus contains truly parallel texts, the correspondent terms can be recognized and linked by means of statistical techniques. The lemmatisation avoids having all the inflected forms scattered. This situation allows for a combination of linguistic and statistical procedures to be used for automatic term extraction. In such a case, a truly multilingual database can be produced, with the same lexicon across languages and for which the terms are all interconnected and correspond to each other.

Scenario no. 2 is represented by a situation where there are pairs of parallel corpora, i.e. different texts in the same domain, but again all parallel to a unique pivot language for which lexical resources and tools are available. This allows us to perform automatic term recognition task with a good success. In such a case, a truly multilingual homogeneous resource cannot be obtained. Instead, the resulting terminological resource is represented by a set of term lexicons in different languages where terminology, with a high probability, will not be the same.

Another scenario is represented by either a truly parallel corpus or a set of parallel corpora which are not linguistically annotated and for which no NLP tools are available. This is the worst solution to be faced, for which statistical procedures of term recognition can only be resorted to, with the risk of producing lists of candidate terms with much noise and where most of the work should be done by hand.

1.1.7 Technical problems

The technical problems fall under two major categories:

- a. identification of suitable texts, and
- b. processing of the acquired texts.

In order to speed up the **identification process** and the extraction of the identified LRs several **tools** (web crawlers, agents) were used for automatically downloading the material needed. However, these tools have not proven to be intelligent enough to avoid extracting resources partly multilingual and thus inappropriate for the INTERA project. As a matter of fact, tools should be enhanced with respect to two directions: improvement of the tools themselves (e.g. more robust alignment techniques) and interoperability of all relevant tools currently used at different phases of processing. The issue of interoperability is closely related with the issue of standards. The promotion and deployment of existing standards as well as the creation of new standards, when these are lacking, is important to ensure viability and re-use of LRs, given the cost of their production.

The task at hand was also impeded by the **lack of formal descriptions** as regards the existing resources and tools for efficiently locating them. More specifically, it has been attested that the description of resources on the basis of **metadata** is crucial for the identification and the proper use of corpora. Numerous surveys have outlined the need for appropriate documentation which would facilitate the identification of the resources on the basis of informative elements (e.g. type/content of resource, encoding format, property rights owner, annotation details, etc.). A combination between formally described LRs and intelligent tools working on the basis of these descriptions would facilitate the resources production and would enhance the services provided by the eContent professionals.

The evaluation of the existing material often revealed the **inappropriateness** of the automatically downloaded resources as regards the **format required**. Any linguistic processing was preceded by the initial task of corpus cleanup, which comprises the conversion of the texts into the appropriate format needed for text processing. The corpus cleanup aims to remove items of the texts which might be necessary for the original mode of publication of the texts (such as pictures, captions, hyperlinks, graphics etc.), but constitute "noise" for text processing in the specific framework.

During the identification phase it had been attested that the resources are available in various formats (html, doc, rtf, pdf, etc.), while in most cases, the processing tools require as input resources in txt format. This incompatibility causes a number of problems as regards the

conformity of the format of the resources to the format required by the linguistic processing tools. Consequently, the conversion of pdf documents into txt format was considered indispensable. However, a few problems had to be overcome during the conversion procedure: the converters available, either free-ware or not, although they performed quite well as regards the English language, have proven not efficient enough as regards the other languages of interest. Post-editing was often needed in order to remedy the sections of the documents that were not properly converted. An important amount of time has been spent on this process, since there exist few converters that can handle non-Latin characters. This point is extremely important in the framework of the INTERA project, since it deals with languages with different character sets. The lack of adequate tools able to handle various non-Latin character sets needs more attention, since it hampers the processing of the less widely available languages and the subsequent building of relevant corpora.

In conformance with the methodology adopted in the project, i.e. of re-using existing material (whenever possible) with the least possible interventions, so as to ensure time and cost efficiency, for the **processing of the texts**, it was decided to re-use only existing tools for each language, without making any modifications to the tools themselves but only conversion(s) of their output. The material delivered, however, at the end of all processes should be conformant to the set standards. The problem faced here was that of the formal harmonization of all the different tagsets used (see also D5.2).

5 Market needs

The Survey of the eContent specialists' needs conducted in the framework of the deliverable D5.1 has been based upon: (a) information available over Internet; (b) the results of previous surveys on the same issue conducted either in the framework of other projects (i.e. ENABLER) or by international organizations (ELDA, LISA, IDC etc.); (c) information acquired through the circulation of a questionnaire that aims to identify the policy of eContent professionals concerning the handling of LRs; (d) the information acquired through the contacts of the involved partners with an important number of actors in the relevant fields.

The eContent professionals addressed by the INTERA project objectives were identified in the deliverable D5.1 and can be classified in the following categories:

- ★ *digital content authors*: this group of users is interpreted in the INTERA framework as the content owners themselves, given that professionals designing and developing applications for electronic publication are included in the second group described below; the needs and technical know-how of this group is too varied and may be more bound to the specificities of the content they wish to make available on electronic media; therefore, they are not *directly* targeted by the project, although they can use the INTERA network as a pool of knowledge for accessing language experts and related information;
- ★ *actors in the Globalisation, Internationalisation, Localization and Translation (GILT) industry*: this group includes professionals offering services for the development of digital content as well as its linguistic translation and cultural adaptation to other languages; LRs are of primary concern to their needs;
- ★ *producers of Human Language Technology (HLT) tools*: providing effective multilingual retrieval and search mechanisms, content management, Machine Translation and translation automation tools is vital to supporting the massive scale of digital content customisation to all languages; HLT integrators have long realized the value of LRs for

their tasks and have been active in the LRs production; in the INTERA project, their needs will be more thoroughly analysed to establish the appropriate guidelines;

- ★ *digital content publishers*: this group includes traditional publishers going digital as well as new actors in the field; given the volume of content they manage, they constitute a primary target for the INTERA objectives;
- ★ *IT/Telecommunications companies*: although presented here as a distinct user group, the activities for which they have been included in the INTERA target group, are shared with other groups: IT/Telecommunications companies, on the one hand, may be developers of HLT tools and, on the other, may act as content distributors (together with Internet services providers, eCommerce retailers and Internet start-ups) although they have no content creating background.

The data that have been collected through this Survey describe the eContent professionals' needs and the market requirements as described hereafter.

5.1 Business practices related to LRs

1.1.8 *Are LRs needed?*

First of all, it is once more confirmed that eContent professionals do need LRs and especially multilingual corpora and terminological ones (see indicatively, Translation Business Practices Report, World Bank Group's Translation Unit (GSDTR, August 2004). Multilingual corpora, especially in the form of parallel and aligned ones, are required to cover a broad range of activities in the eContent world.

1.1.9 *How are LRs acquired?*

Given this need the eContent professionals either employ language specialists in order to develop them in-house or they buy LRs from distribution agencies (i.e. ELDA). It is also attested that there are few multilingual LRs freely available and downloadable over the Internet, which makes the role of LRs producers and distribution agencies more crucial. As regards the terminological resources it is attested that the communication of specialist knowledge and information is bound with the creation and dissemination of terminological lexicons. The quality of communication depends to a large extent on the quality of terminology employed. It should be borne in mind that communication is not solely monolingual especially not within Europe. In fact, there is a clear trend towards an increased awareness of multilingual issues, despite the predominance or at least the lead function of English in the technical, business, economic, political and cultural fields. Readily accessible, up-to-date terminology will play an important role in multilingual information society in the 21st century.

The identification of this demand reinforces the assumption that there is certainly a market for multilingual resources, if we by this mean that there are people willing to sell resources and there are people willing to buy. But it is very hard to know anything on more specific market characteristics. This is because marketing and distributing LRs is a relatively new and unexploited activity. Without disregarding this fact, we will make an attempt to define the properties of the LRs needed by the eContent professionals so as to take them into account to the proposed methodology concerning the LRs production.

1.1.10 *Which languages are mostly needed?*

Surveys show that English is currently the most represented language on the Internet. However, the same surveys show that there is a growing tendency to include other languages as well. These include not only the other competing "international" ones but, most important, even the less spoken national languages. It is becoming increasingly evident that people prefer to communicate in their own languages.

According to the Language Resource User needs and market analysis conducted by ELRA in 2000 (<http://www.elra.info/>), English, French, German, Italian and Spanish are currently the most desired languages for LRs. According to the Translation Business Practices Report, the language combinations requested most are English into Chinese, German, Portuguese, Italian, Scandinavian languages, Japanese, French, Spanish, Korean, Dutch, Greek, Indonesian, Vietnamese, Slovenian, Ukrainian, Serbian and Croatian.

In addition, past initiatives and projects of European or international organizations (e.g. ElsNet goes East, TRACTOR, etc.) confirm a general interest in Balkan and eastern European languages. Moreover, the upcoming expansion of the European Union with more candidate countries, together with its policy of equality between languages and protection of cultural (including linguistic) heritage, makes the problem of LRs production even more intense. This situation forces the eContent specialists to create bilingual, trilingual or multilingual sites and move across a variety of languages, in order to reach as large an audience as possible.

Following this situation the majority of companies tend to provide their content in at least 2 languages (corresponding most likely to a home market and another market). It is also observed that a considerably high percentage of eContent professionals tend to provide their content in more than 5 languages (LISA 2003). Looking ahead eContent players are planning to increase the number of languages over what they now provide (Figure 2). Since localization is a precondition of doing effective international business in many markets, this trend indicated that international markets are seen as *multilingual* markets.

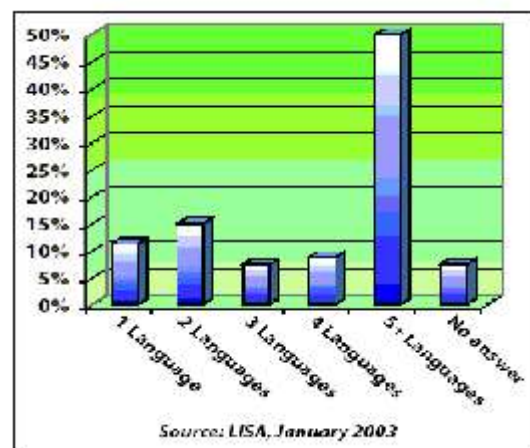


Figure 2. Number of languages in which content is planned or desired

These tendencies should be considered in terms of the difficulty in identifying qualified translators available especially for the less common languages. This problem proves the intensive need for multilingual and terminological resources in a variety of languages which makes the role of LRs producers crucial and the modernisation of the LRs production process indispensable.

1.1.11 Raw or processed data?

Surveys also attested the need for multilingual resources *ready* to be processed by relevant tools or already processed to a certain level. More specifically eContent professionals need

LRs that can be exploited with the minimum of adaptation. This observation renders *alignment* the main identified requisite. In general, alignment refers to the establishment of links between two pieces of multilingual texts, where one is the translation of the other, whatever the granularity level. A rough classification of the proposed alignment procedures is based on the definition of the text unit, or the translation unit in this case. Unsurprisingly, the initial research experiments focused on sentences, partly due to their major significance as text units, but also due to the high degree of translational ambiguity compaction that the use of sentences as translation units entails. In the last years, however, interest for establishing equivalences at levels lower than the sentence, such as words, single-word and multi-word terms or even clauses, has grown relatively high.

1.1.12 Metadata descriptions

Furthermore, crucial to the proper use of corpora (as well as of any language resource) is the *description* of the resource; numerous surveys have outlined the need for appropriate documentation which would facilitate the identification of the resources on the basis of informative elements (e.g. encoding format, property rights owner, annotation details, etc.). The description of the resources available is quite useful since it helps producers to:

- locate useful language resources within a distributed domain of resource metadata descriptions stored and offered at several data centres in Europe;
- immediately and transparently start useful operations on the located resources.

1.1.13 Are standards crucial?

eContent specialists claim that data exchange standards are a specific concern to the reusability of data. The majority of the target group notes that the LR_s externally acquired usually need further processing, at least conversion of their format. As regards the standards used, most of the companies seem to trust XML and SGML standards. The high level of XML use also points to the maturation of tools for dealing XML.

It is also attested (Maegaard, B., Khalid Choukri, Valérie Mapelli, Mahtab Nikkhou & Claus Povlsen (2003), Language resources - Industrial needs, ENABLER Deliverable 4.2) that "several companies have a positive attitude towards standards, but as long as there is more than one standard, they do not feel the advantage". Others claim that the standards available are not adequate for their applications. For the companies, the most important thing is that data has a good quality and that the description is formalised. This makes conversion possible. Compatibility of the LR_s with the companies' own tools or the tools available by HLT producers is also taken into account.

As regards the formats translation memories are stored in, the majority of companies tend to use either commercial formats or the open standard TMX (Translation Memory exchange). This situation is also attributed to an increasing client request for standards (LISA 2002). Moreover, companies aim at ensuring reusability and easy integration of additional technologies. It should be noted that although companies are willing to use international standards in order to make their content reusable and exchangeable they claim that they are not sufficiently educated as regards the formats and the relevant tools.

1.1.14 Texts without processing tools?

Apart from the multilingual resources needed it should be taken into account the request for tools appropriate either to easily extract and create the LR_s or to effectively process them. Among the main tools needed for the production of multilingual LR_s are the translation memories. According to the surveys TM technologies are rapidly maturing and most of the companies use TMs for everything they can and are looking at extending their usage. A

considerable percentage of eContent professionals report using TMs for purposes beyond simple translation (quality assurance, term replacement etc.)

TMs could also serve as a basis for the development of a variety of commercial applications such as multi-/ cross- lingual information retrieval systems, automatic translation systems, multilingual summary systems etc. The aforementioned applications are among the predominant ones as regards the eContent professionals according to the surveys.

As regards the terminological tools a demand for terminology extraction, terminology consistency checking and statistics reporting tools is also attested.

It should also borne in mind that language processing tools are not well integrated and interoperable. Terminology databases, translation memory systems, machine translation systems often lack seamless integration that would increase productivity among translators and terminologists.

1.1.15 Other issues

Apart from the aforementioned needs of eContent professionals there are also other issues that should be discussed concerning the acquiring policies and the pricing practices of the LRs. These issues should be borne in mind since they are strongly related to market needs and thus serve as a basis for the model proposed in the deliverable.

Various surveys have attested the difficulties related to the acquisition of the multilingual resources and the terminology data needed. This difficulty is strongly related to availability problems already discussed in the deliverables D5.1 and D5.2 of the INTERA project. One of the main obstacles in the acquisition of the LRs is the overall policy concerning their dissemination which is dictated from the considerable high production cost. This triggers a kind of vicious circle, on the basis of which developers are discouraged to invest, and funding agencies to support infrastructural endeavours. Another constraint is that national activities are directed towards the production of resources in their national languages. The field of multilingual resources production suffers an unavoidable condition of disadvantage and consequently the eContent world, where this kind of resources is crucial, is handicapped. In addition, the lack of extensive dissemination and publicity from the producers' side makes the identification of LRs harder.

According to the ELRA survey conducted in 1997 (M. Nilsson, ELRA MARKET STUDIES IN 1997, <http://www.elra.info/>), when it comes to the basis for the decision to buy LRs, the factors taken into account are: cost effectiveness, time perspective for production and costs, suitability for the task, ownership/user rights, availability, quality, reusability, customisability and monitorability.

As regards the availability of industry resources, many companies are protective of their data and prohibit its distribution. As a consequence most of the translators and the localizers struggle to find the suitable multilingual and terminological resources in order to achieve terminology consistency in an environment where there is a reluctance to share terminology. This situation puts further burdens on the creation process of LRs and on distribution and dissemination processes as well, since sharing LRs across industry in a profitable way would increase consistency and reduces localization effort.

6 Business model for the LR production community

6.2 Business model: definitions

Speaking in strictly business terms, a business model in its most basic sense is a tool for setting up the methodology by which a company can sustain itself - that is, generate revenue by specifying where it is positioned in the value chain. To this end, it should contain a description of the operations of the business including the components and the functions of the business, along with the revenues and expenses that this business generates.

A business model is different from a business plan in that the latter is a more detailed document “setting out the objectives of a business over a stated period (often three, five or ten years)” and is usually intended for new businesses. However, like business plans, a business model should ideally contain quantitative figures on expected costs and profits involved in the business cycle. Being a mechanism by which a business intends to generate revenue and profits, the business model is a summary of how a company plans to serve its customers and promote its products or services and involves both strategy and implementation. It should provide for:

- customers selection (market segment targeting)
- definition of the company products and offerings that differentiate the company from its competitors
- creation of utility for its customers
- methods for acquiring and keeping customers
- promotion and distribution strategy
- definition of the tasks to be performed
- configuration of its resources
- capturing of profit.

There are various types of business models, depending on the nature of the business addressed. Generally, the business models of service firms are more complex than those of manufacturers and resellers. The oldest and most basic business model is the shopkeeper model. This involves setting up a store in a location where potential customers are likely to be and displaying a product or service. Over the years, business models have become much more sophisticated. The *bait and hook* business model (also referred to as the razor and blades business model or the tied products model) was introduced in the early 20th century. This involves offering a basic product at a very low cost, often at a loss (the "bait"), then charging excessive amounts for refills or associated products or services (the "hook"). Examples include: razor (bait) and blades (hook); cell phones (bait) and air time (hook); computer printers (bait) and ink cartridge refills (hook); and cameras (bait) and prints (hook). An interesting variant of this model is a software developer that gives away its word processor reader for free but charges several hundred dollars for its word processor writer.

But times are changing and companies must continuously rethink their business design. Companies must change their business models as value migrates from industry to industry. Ultimately the success or failure of a company depends first on how well its business design matches their customers' priorities. However, the literature shows that many companies find it difficult to manage innovations that fall outside their previous experience, where their earlier beliefs and practices do not apply.

6.3 LR production: the business

To define, therefore, a model for a given discipline is vital, yet not always a trivial task, since there are many parameters to be taken into consideration. Among these, the identification of

the nature and peculiarities of the business at hand is the first. LR production is an endeavour undertaken in a wide range of organisations that fall to one of the following categories:

- (a) academic units,
- (b) research centres and institutes, and
- (c) companies that run on their own funds.

On the other hand, certain companies, such as software developers or eContent professionals, have departments whose mission is the production of LRs tailored to their own needs. The latter either follow up-to-date and widely accepted standards, or develop their own ones.

Given the diversity of the types of LRs producers, which entails entirely different perspectives, sets of beliefs and practices, and, obviously, market and marketing experience in the field of LRs production, the definition of a global, appropriate for all, business model seems almost impossible. Given also the nature of the product in question, namely, **language** resources, this task is furthermore hampered by the diversity in the existing (or not) infrastructure, but also in cultural approaches related to the issue of language.

LRs are costly to produce and maintain, and - in most cases - they come as by-product in the process of language engineering applications development and evaluation. This has an impact on the characteristics of application-specific LRs, that are not always portable to applications other than the ones they were initially built for. Re-usability of LRs, therefore, is an issue that has to be taken into consideration in a proposed model.

Another peculiarity of the LRs business is that they can be produced on-demand, or constructed, as the result of R&D projects, with no specific plan for their marketing thereafter.

6.4 The business model

The business model proposed here is intended for professionals that are involved in the production of multilingual LRs such as corpora, lexica, terminological databases, etc. As discussed above, there are inherent and clear cut distinctions in the processes and functions of organisations involved in the task of LR production (their being academic institutions, research centres, or private companies and firms). In the proposed model there is no explicit distinction in the business practices that better suit each; the union rather than the intersection of their respective business models is hereby presented. The model presented here addresses the following issues:

- (a) Market segment targeting
- (b) LR production methodologies
- (c) Resource configuration
- (d) LR portability and re-usability
- (e) LR promotion and distribution strategy

It should be noted, however, that due to the different level of complexity and the diversity of the processes entailed by the legal as well as the financial status of each of the aforementioned types of organisational bodies, no quantitative estimates with respect to costs and expected profits will be provided. This is also due to issues related to the fact that the costs of LRs production are very variable and precise figures could not be provided, even by experts in the field of LR dissemination.

Moreover, the INTERA experience proved that LR development costs depend on a variety of factors, such as languages (widely used vs. almost exotic), types of annotations, infrastructures available or not (processing tools, LRs producers' experience and know-how), IPR clearance costs, etc. Local market conditions - where subcontracting is involved - allow for a relatively restricted development cost which combined with a limited offer could potentially lead to a substantially net margin profit.

1.1.16 Market-oriented development of LRs

Market targeting is the first issue to be considered by companies of any kind. To be efficient in their business, LR producers should be able to define the user groups their products are targeted at (i.e., professionals that are expected to be interested in the resources to be constructed) and outline their respective needs and requirements. To this end, LR producers should conduct surveys or consult the results of already existing ones in order to be aware of market trends with respect to types of LRs, domains, languages, and relevant metadata that possible customers would need in their businesses. The identification of user needs and requirements on the basis of the professionals' working habits and processes was achieved within the framework of INTERA by exploiting the results of a number of previous initiatives to roadmap the state-of-the-art in multilingual LRs, in combination with new initiatives undertaken in the framework of the project and targeted to the eContent world. The surveys conducted in the framework of the ENABLER project (Maegaard et al. 2003, Gavrilidou & Desipri 2003) provided insights as to the existence and availability of different types of LRs, language demand, domains of interest, standards, etc. Other surveys, such as those conducted by ELRA and its distribution agency ELDA aiming at determining the needs of users with respect to available and potentially available LRs (<http://www.elra.info/>), or surveys available over the Internet through the sites of international organizations such as LISA and IDC or consultancy firms (<http://www.globalsight.com>, LISA 2001, LISA/AIIM 2001, LISA/OSCAR 2003) shed a light as to the availability of resources and relevant tools.

The market-oriented resource production perspective is expected to have a direct impact on LR specifications with respect to a number of characteristics. More precisely, it will:

- determine the domains that are of interest to potential users, so that domain-specific resources are favoured.
- specify the languages or language pairs most needed that will grow revenues for the LR producer
- define the corpus processing specifications, i.e., alignment, external and internal structural annotation, and linguistic annotation.

However, **realistic considerations** are the other side of the coin. LR producers should anticipate possible shortcomings in all stages of production. When listing the resource requisites and content specifications on the basis of documented users' needs, the "best scenario" is described, i.e. what ideally would answer the user needs. But, while proposing a realistic production model, different scenarios should be accommodated according to different likely situations that could arise, especially for under-represented languages. The quality of the final resources produced is strictly intertwined with the existing source material, its (re)usability, the presence of linguistic annotation or the availability of tools to perform linguistic analysis, the compatibility/reusability of different linguistic analyses, etc. Data availability, representativeness, and size, together with availability of language processing tools are crucial factors to be taken into account in a very *realistic production model*, since they are variables with strong repercussions on the task and the process of multilingual LR production.

1.1.17 LR production methodology

A unique production methodology for LRs which can be applied to all cases and uses cannot be defined on sound basis, since the production methodology is heavily influenced by the kind of data to be used and the final purposes for which a resource (corpus, lexicon, terminology, etc) is produced. In Deliverable 5.1 it was already argued that the requirements to be fulfilled by a resource strongly depend on the final use of the resource itself.

Despite certain obvious drawbacks it is indisputable that World Wide Web is a mine of language data of unprecedented richness and ease of access (Kilgarriff and Grefenstette 2003). It is obviously the largest source of "available" language resources. LR producers should take advantage of the web in order to modernize either the development of LRs from scratch or the enrichment and the updating procedures so as to ensure the reusability of old ones. Moreover the development of language engineering applications usually involves the creation of specific corpora. In order to be efficient, LR producers should investigate the possibility of using existing material. In this context, there is a great need for usage of efficient methods to collect data and to employ methodologies that are up-to-date. This methods in order to be considered as attractive solutions should lead to the automatization of the LR production task.

In order to speed up the identification process and the extraction of the identified data suitable for the building of LRs several tools (web crawlers, agents) are currently being used for automatically downloading the material needed. In addition, several techniques have been employed aiming at exploiting the web content. The idea of building a corpus using automated search engine queries originates from Ghani et al. (2001), who apply it to the creation of minority language corpora. Several techniques have followed Ghani's idea. Indicatively we mention the following:

- STRAND Web-Mining Technique which is exploiting an observation about the way web authors disseminate information in multiple languages.
- Parallel Text Miner (PTMiner) exploiting already existing web search engines to locate pages by querying for pages in a given language that contain links to pages that are likely to be in the other language of interest
- Bilingual Internet Text Search (BITS) which starts with a given list of domains to search for parallel texts.

Such techniques, although relatively new and still not sufficiently tested and explored, should be taken into account when it comes for LR production. However, it should be borne in mind that, especially for the creation of parallel multilingual resources, which are of great interest,

they would probably need to be improved in terms of interoperability of relevant tools currently used at different phases of processing. On the other hand when dealing with the building of domain specific LRs content based information should be taken into account.

As regards the collection of comparable corpora, which seem to be a viable solution concerning the terminology extraction, automatic techniques of acquiring the appropriate data should be quite “intelligent”. This means that crawlers and agents employed should probably incorporate lexical information in order to identify the resources required.

In what concerns terminological resources, Auger wrote in 1989: “In a futuristic scenario, terminologists will have access to huge data (or knowledge) banks; from these banks, they will download items belonging to their corpora; they will go through these items without having to previously tackle the text manually; they will automatically establish their working terminologies, contrasting the incoming terms with the semantic descriptors that will later be used in the automatic writing of definitions; they will classify, choose, merge and edit data bases reducing their intervention to a minimum. Their work station (...) equipped with advanced and intelligent office tools, will allow them to control, by themselves and throughout the whole process, the elaboration of their product and carry it under the best conditions”.

Fifteen years later, we can see that the introduction of computer tools has heavily changed the methodology of terminology production, allowing for fast collection of amounts of data which could not be harvested by hand. However, the availability of computerized tools and of automatic or semi-automatic procedures for term extraction has highlighted the existence of other kinds of problems and shortcoming that had not been previously identified.

The sheer availability of data in electronic format is not a sufficient condition for the production of good-quality terminologies.

1.1.18 Text processing tools

In order to enhance the LRs production effort, the re-use of existing tools is considered crucial. It is true that an increasing number of tools are available for text processing; however, most of the language dependent tools are oriented mainly towards the major languages. Moreover, information concerning the existence, availability and operation of existing tools is not easy to locate. This gap is addressed by the other pillar of the INTERA project, which tries to remedy it through the building of an integrated European Language Resources area, whereby the users can locate resources and suitable tools and actually run the tools on identified appropriate resources remotely. Additionally, tools must be enhanced with respect to two directions: improvement of the tools themselves (e.g. more robust alignment techniques) and interoperability of all relevant tools currently used at different phases of processing.

1.1.19 Resource configuration

The re-use of already existing resources (either raw texts or already processed and annotated material) is considered indispensable for the quick and efficient production of new LRs.

The usage of the web as a source for the creation of LRs is imperative. However, when the aim is the production of multilingual resources, the usability of the web is not so simple and straightforward procedure. The multilinguality of the web has been discussed in previous section.

As regards IPR issues, it should be noted that many providers of web-based content retain their copyright. This means that not all web-accessible data can be freely used and shared, but that copyright notices have to be taken into account, permissions sought and the necessary agreements signed.

1.1.20 LR portability and re-usability

The issue of interoperability is closely related with the issue of standards. The promotion and deployment of existing standards as well as the creation of new standards, when these are lacking, is important to ensure viability and re-use of LRs, given the cost of their production.

1.1.21 Dissemination and promotion strategies that could apply to LR producers

Use of existing distribution channels

Major distribution agencies such as ELDA and LDC facilitate business transactions between franchised distributors and their trading partners. They maintain a catalogue operation that connects a large number of product manufacturers with volume and retail buyers.

Subscription model

The subscription business model is a business model that has long been used by magazines and record clubs, but the application of this model is spreading. Rather than sell products directly, more and more companies are selling monthly or yearly access to a product or service. This, in effect, converts a one-time sale of a product into a recurring sale of a service.

In addition one can have a subscription on a discount pass. Businesses benefit because they are assured a constant revenue stream. This greatly reduces uncertainty and the riskiness of the enterprise. Also, in many cases (such as integrated software solutions), the subscription pricing structure is designed so that the revenue stream from the recurring subscriptions is considerably greater than the revenue from simple one-time purchases. In some subscription schemes (like magazines), it also increases sales, by not giving subscribers the option to accept or reject any specific issue. This reduces customer acquisition costs, and allows personalized marketing or database marketing. Subscription business models also have their drawbacks. The business must commit to a large infrastructure to manage and track subscriptions.

Applying these principles to the LRs production business, the producers could make their LRs accessible over the Internet via a service operated on a subscription base. This service could provide also the tools for specific operations on the texts, or allow via an interface the presentation of the results of specific operations. This service could offer

- updates to users or customers
- by-products (tools, terminologies, dictionaries, sub-corpora...)
- open access to corpora, with a charge for other products option
- on-line use of LRs
- remote usage of corpora and tools (charge service).

Dissemination via the Internet

Usage of mailing lists, banners in selected sites relevant to LRs, eContent and HLT communities are among the most easy and affordable methods for dissemination of produced LRs.

7 Conclusions

The present report tried to outline methods and practices which would make the production of LRs attractive to the LRs producers. Given the diversity of the field and of various parameters that decide the implementation of such an endeavour, quantification of the practices discussed in relation to cost and revenue was not a feasible task.

At a general level, the production methodology is heavily influenced by the following factors:

- *Lack of integration among computer tools working at different levels of analysis:* different tools may require different formats for the data to be processed. Format conversion usually is not straightforward and can lead to loss of information. To this end, adherence to standards in character encoding and/or use of de-facto standards such as XML is strongly recommended.
- *Lack of compatibility among the resources themselves:* especially from the point of view of the production of multilingual resources, compatibility among the resources to be processed is of utmost importance. This means, for instance, not only enforcing compatibility in data encoding and representation, but also ensuring that the resources are compatible from the point of view of the additional, linguistic and non-linguistic information which is added to the raw data. For instance, it might be the case that different resources adhere to different conventions in the representation of linguistic information, both at the level of how linguistic information is described (the breadth of part-of-speech tags, for instance), and at the level of how much linguistic information is provided (i.e. the depth of analysis expressed by the annotation system). Once again, compliance with agreed-upon standards is recommended, as well as harmonization among the different tagsets used in the various resources. Ideally, all resources should use the same convention of linguistic annotation; when this is not possible, it is recommended that a harmonized tagset is used, or that conversion procedures from the proprietary tagset to a common, standardized one is provided.
- *The limited number of existing corpora, especially in languages other than English:* as already extensively argued, the bottleneck represented by scarce availability of language resources, especially for less-represented languages, is the factor that most harms the entire process of LRs production. Especially in the case of terminology, the few language resources available result in terminologies that suffer from deficiency in quantity and quality of the terms produced, especially when automatic, statistic-based procedures are used. The scarce availability of corpora for non-English languages is also the reason why it is difficult, when not of little use, to draw detailed recommendations, since these heavily depend on the particular configuration of input resources that are available in the single, distinct cases. Hence,
- *The particular configuration of resources available:* the particular methodology to be adopted for the production of multilingual terminological LRs must be carefully adjusted to the idiosyncratic situation to be handled, where by situation we mean the type of languages, the quantity and quality of resources, and the purposes for which the resource is being built. This is what we have defined as a *scenario*.

In conclusion, the experience gained during the INTERA project calls for more resources, of good quality, and compliant with sound standards. Lesser favoured

languages can benefit from building parallel resources where one language is represented by English.

Another general recommendation is that a criterion of practical feasibility be followed, balancing the constraints imposed by corpus size, languages, and users' and standards requirements. This is deemed the only viable and reasonable solution, especially from the point of view of prospective users that will have to apply the model that is the outcome of the INTERA project.

In conclusion, the experience gained during the INTERA project calls for more resources, of good quality, and compliant with sound standards. Less favoured languages can benefit from building parallel resources in combination with English.

Another general recommendation is that a criterion of practical feasibility be followed, balancing the constraints imposed by corpus size, languages, and users' and standards requirements. This is deemed the only viable and reasonable solution, especially from the point of view of prospective users that will have to apply the model that is the outcome of the INTERA project.

What the business model, then, proposes to the LRs producers is:

- (a) production of LRs with the market in mind
- (b) if the original goal is not related to the market needs, try to be as close to market needs as possible, in order to transform the produced LRs into attractive resources for the market with minimal effort
- (c) careful survey of the situation, as regards existing LRS, tools, methods, similar endeavours that could be useful (with a realistic perspective)
- (d) flexibility in goals, such that would allow for possible need of change of plan according to the results of (c)
- (e) open mind for new advances in technological solutions that would facilitate their work
- (f) reusability of resources and tools when and where possible
- (g) therefore, adherence to existing or emerging standards, in order to ensure interoperability, automatization of processes, updates and enrichment of resources.
- (h) usage of metadata for the description of the resources, so that they can be easily identifiable by prospective users
- (i) investment in dissemination and promotion.

8 References

- ★ Allen, J., and K. Choukri. 2000. Survey of Language Engineering needs: a Language Resources perspective. LREC 2000.
- ★ Auger, P. 1989. "La terminotique et les industries de la langue". *Meta* 34,3. 450-6.
- ★ Baroni M. and S. Bernardini. 2004. BootCaT: Bootstrapping Corpora and Terms from the Web. LREC 2004.

- ★ Chesbrough, H. and R. S. Rosenbloom. 2002. The role of the business model in capturing value from innovation: evidence from Xerox Corporation's technology spin-off companies, *Industrial and Corporate Change*, vol. 11, Number 3, 529-555.
- ★ Choukri, K. 2000. ELDA LE4-8335 Final Report.
- ★ Gavrilidou M., Desipri E., Labropoulou, Piperidis S., Calzolari N., Monachini M., Soria C. 2003. Technical Specifications for the Selection and Encoding of Multilingual Resources, INTERA D5.1.
- ★ Gavrilidou, M. E. Desipri, P. Labropoulou, S. Piperidis, N. Calzolari, M. Monachini & C. Soria. 2004. Technical specifications for the selection and encoding of multilingual resources, INTERA (Integrated European language data Repository Area), Deliverable 5.1.
- ★ Gavrilidou, M., E. Desipri. 2003. Final Version of the Survey, ENABLER Deliverable 2.1.
- ★ Ghani, R., Jones, R. and Mladenec D. 2001. Mining the Web to create Minority Language Corpora, *CIKM 2001*, 279-286.
- ★ Johnson, G., K. Scholes, 1993. *Exploring Corporate Strategy*. Prentice Hall Int.
- ★ Kilgariff A. and Grefenstette 2003. Introduction to the Special Issue on the Web as a Corpus, *Computational Linguistics*, vol. 29, Number 3.
- ★ LISA. 2001. *The LISA Globalization Strategies Awareness Survey*.
- ★ LISA/AIIM. 2001. *The Black Hole in the Internet: LISA/AIIM Globalization Survey*.
- ★ Maegaard, B., Khalid Choukri, Valérie Mapelli, Mahtab Nikkhou & Claus Povlsen. 2003. Language resources - Industrial needs, ENABLER Deliverable 4.2.
- ★ Nilsson, M., 1998. ELRA MARKET STUDIES IN 1997, <http://www.elra.info/>.
- ★ Resnik Ph., and N. Smith. 2003. The Web as a Parallel Corpus, *Computational Linguistics*, vol. 29, Number 3.
- ★ The Oxford Dictionary of the Business World.
- ★ World Bank Group's Translation Unit. GSDTR. , Translation Business Practices Report, August 2004.