Language Resources Production Models:

the Case of the INTERA Multilingual Corpus and Terminology

Maria Gavrilidou[1], Penny Labropoulou[1], Stelios Piperidis[1], Voula Giouli[1], Nicoletta Calzolari[2], Monica Monachini[2], Claudia Soria[2], Khalid Choukri[3]

[1]ILSP/IRIS, Athens, Greece
{maria, penny, spip, voula}@ilsp.gr
[2]CNR-ILC, Pisa, Italy
{nicoletta.calzolari, monica.monachini, claudia.soria}@ilc.cnr.it
[3]ELDA, Paris, France
{choukri@elda.fr}

Abstract

This paper reports on the ***multilingual Language Resources*** (MLRs), i.e. parallel corpora and terminological lexicons for ***less widely digitally available languages***, that have been developed in the INTERA project and the methodology adopted for their production. Special emphasis is given to the reality factors that have influenced the MLRs development approach and their final constitution. Building on the experience gained in the project, a ***production model*** has been elaborated, suggesting ways and techniques that can be exploited in order to improve LRs production taking into account realistic issues.

1. Introduction

This paper reports on the results of the INTERA project (Integrated European language data Repository Area, http://www.elda.org/intera) an eContent program that had a twofold task:

- to build an ***integrated European Language Resources (LRs) area***, and
- to ***produce new multilingual LRs (MLRs)***.

The paper focuses on the second axis of the project, presenting the MLRs produced, namely ***multilingual parallel corpora*** and ***terminologies*** extracted from these, the methodology used for their production and the experience gained. Finally, it presents a MLRs production model elaborated on the basis of this experience, which is up-to-date, compliant with existing standards but also viable and attractive for digital content producers.

2. Resources description

INTERA aimed at the production of MLRs for the ***eContent business actors***, namely ***parallel corpora*** and the related ***terminological lexica***. Moreover, these resources should be developed for the "***less widely available languages in the digital world***[1]". This notion refers to languages which suffer from poor representation in the LRs field, scarcity of raw material in digital form, and, often enough, lack of robust processing tools that would allow their quick integration into Human Language Technology (HLT) applications and/or easy exploitation in the eContent industry (Gavrilidou et al., 2003; 2005). This term has been preferred over other terms such as "less spoken", "less used", "minor" languages etc., which, not only have a pejorative aspect (all languages are equally important!), but also fail to capture the real situation: many languages of this category, such as eastern European and Balkan ones, actually appear among the forty most widely spoken languages all over the world (http://www.globallanguages. com), while they appear in few, if at all, catalogues of digital resources.

[1] Other terms currently employed for the same notion are: *resource-poor* and *resource- scarce languages.*

2.1. Parallel corpora

2.1.1. Size and domain coverage

The INTERA project aimed at the construction of a multilingual parallel corpus of ***12 million words*** (MWs) in several languages, mainly from the Balkan area. However, this goal had to be re-formulated due to availability problems (see section 3.1.3). As a consequence, the final text collection is a ***multilingual comparable corpus***, made up of ***bilingual parallel sub-corpora***: instead of having the same texts in all languages, pairs of parallel text collections have been produced, belonging to the same domains (Table 1). English (EN), the pivot language, always represents one member of the pair, while the other is Bulgarian (BG), Greek (EL), Serbian (SR) or Slovene (SL). 2 MWs have been produced for each of the BG-EN and SR-EN pairs and 4 MWs for each of the EL-EN and SL-EN ones, reaching thus the intended size of 12 MWs.

Domain	**Language pair**			
	BG-EN	GR-EN	SR - EN	SL - EN
Law	■	■	■	■
Health		■	■	
Education	■	■	■	
Tourism		■		
Environment		■		
Politics	■			
Law - Politics	■			
Finance	■			

Table 1: Domain coverage per language pair

2.1.2. Text processing

All texts are

- aligned at sentence level (formatting conformant to the TMX standard, http://www.lisa.org/tmx/),

- structurally annotated at sentence level (external annotation adheres to the IMDI metadata schema (http://www.mpi.nl/world/ISLE/schemas/schemasframe.html) and internal annotation to the XCES standard (http://www.cs.vassar.edu/XCES/), i.e. the XML version of the Corpus Encoding Standard (CES, http://www.cs.vassar.edu/CES/CES10.html)),
- morphologically (below-PoS) tagged and lemmatized (formatting conformant to the XCES standard which incorporates the EAGLES guidelines for annotation (http://www.ilc.cnr.it/EAGLES96/home.htl).

Text processing at all levels has been validated *automatically* as regards *form* (conformance to the specifications), whereas *content* has been *manually validated by native speakers* at the level of *alignment*.

2.2. Terminologies

The INTERA terminological collection amounts to a total of ***17357*** terms, unevenly distributed across the five project languages. Due to the particular configuration of the corpora representing the basis of the extraction process, it is more appropriate to talk about multiple multilingual terminologies, one for each domain. The overall archive is organized into eight packages, each one corresponding to a particular domain (Table 2).

	EL	BG	SR	SL	EN	Sum dom.
Law	1232	279	1436	2052	5042	**10041**
Law-Politics		426			424	**850**
Politics		39			39	**78**
Education	1707	81	232		1679	**3699**
Environment	182				166	**348**
Health	518		201		604	**1323**
Tourism	524				480	**1004**
Finance			14		14	**28**
Sum lang.	**4163**	**825**	**1883**	**2052**	**8448**	**17357**

Table 2: Term distribution over domains & languages

The terminological entries were encoded according to the TMF standard (Terminological Markup Framework, ISO 16642/2001, http://www.loria.fr/projets/TMF). The following types of information were selected for encoding, under the form of specialized Data Categories:
Id: unique identifier of the entry.
Domain: the field of special knowledge for each entry.
Language: a unique identifier of the language of the entry.
Term: a designation of a general concept in a specific subject field.
Grammatical info: a category assigned to a word based on its grammatical and semantic properties.
Context: the URI where the lemma is attested.
Component: for the encoding of the components of a multiword term entry.
Rank: for the position of the components inside the term.

3. Resources production methodology

3.1. Parallel corpora

3.1.1. Related efforts

Despite the well-acknowledged usefulness of LRs for the advancement of HLT, as well as for the enhancement of theoretical and applied linguistics research, very few languages benefit from the existence of such resources of adequate size and quality. For instance, the ENABLER project has identified an important gap in the LRs (even monolingual) production as regards Balkan languages (Desipri et al., 2003). As for multilingual and parallel resources, these constitute a rare commodity:

- most MLRs are mainly comparable corpora, i.e. made up of texts in various languages, not translational equivalents of each other, but similar in size and domain (e.g. MULTEXT-EAST for East European languages, http://nl.ijs.si/ME/ & Erjavec, 2004);
- parallel textual resources are composed mainly of bilingual texts in the "resource-affluent" languages, i.e. English, French, German, and, recently, Arabic and Chinese (Mihalcea & Simard, 2005)[2].

INTERA aimed to remedy this gap in the languages selected, aiming not only at the production of such resources but, more importantly, at the development of a production model that would lead to the proliferation of LRs construction initiatives even for languages usually considered as non-profitable.

3.1.2. Initial methodology for text collection and processing

One of the parameters for the design of LRs production is that of the intended users, i.e. the identification of ***user needs and requirements***. In the context of INTERA, this referred mainly to the eContent business actors. To this end, the results of a number of previous initiatives to roadmap the state-of-the-art in MLRs[3], in combination with new initiatives undertaken in the framework of the project, targeted to the eContent world (Gavrilidou et al., 2003), have been exploited.

On the basis of the results of these studies and taking into consideration the intended application of *automatic terminology extraction*, the specifications regarding the INTERA MLRs have been defined as follows:

- *domains*: eContent users are more interested in special rather than general language, with health/medicine, tourism, education, law, automotive industry and tele-communications being the prevailing

[2] Currently there is a growing interest in the production of LRs for resource-poor languages, among which one initiative (BABYLON, Natural Language Processing for Languages with Scarce Resources, http://mira.csci.unt.edu/~babylon/) includes activities on building bilingual corpora for them.

[3] These include surveys conducted in the framework of the the ENABLER project (Maegaard et al., 2003; Gavrilidou & Desipri, 2003; Calzolari et al., 2004), surveys carried out by organizations involved in LRs activities such as ELRA (http://www.elra.info/) and LISA (LISA, 2001; LISA/AIIM, 2001; LISA/OSCAR, 2003) and big consultation firms in the Information Technology domain, such as IDC (http://www.idc.com/) and Globalsight (http://www.globalsight.com) that have carried out studies related to LRs uses.

domains. Among these, we focused on the domains of *tourism, health, education, law* and *finance*, which are related to predominant digital activities (eTourism, eHealth, eLearning, eGovernment and eCommerce).

- *languages*: the focus being on *less widely digitally available languages*, the selected set was Bulgarian, Greek, Serbian and Slovene, combined with English.
- *standards*: according to the surveys, the use of standards is appreciated by eContent professionals, as it permits reusability and interoperability; thus, adherence to international standards for text processing and annotation, and terminology encoding was considered crucial.

3.1.3. Facing reality and its consequences

The process of identification of existing material in the languages of interest was challenged by problems at all phases concerning almost all the objectives. Thus:

- It unveiled the real status of the web, which is attested to be **multilingual but not parallel**[4]: parallel texts in multiple languages are extremely rare, especially for the less widely available ones, given that international organizations and multilingual portals usually (and understandably) provide their content mainly in the most dominant languages.
- **Lack of "true parallelness" of the web** has also been attested: thorough checking of seemingly parallel texts revealed that only parts of them were indeed parallel, while, quite often, "translations" proved to be summaries or paraphrases of the original text.
- The identification of existing LRs in the **predefined domains** was not without problems as well.
- The availability issue was further hampered in several cases by **IPR problems**[5]. It should be noted that, despite the general feeling that web texts can be freely reproduced, especially for research purposes, this is not true, a fact that ruled out automatic text acquisition processes such as mining the web.
- The task at hand was also impeded by the **lack of formal LRs descriptions** and **intelligent tools** for their efficient identification. As pointed out by numerous surveys, appropriate documentation facilitates the identification of the resources on the basis of informative metadata elements (e.g. type/content of resource, encoding format, property rights owner, annotation details etc.). A combination between formally described LRs and intelligent tools exploiting these descriptions would facilitate resource production and would enhance the services provided to the intended LRs users.
- Finally, the task of text processing was hampered by the **scarcity of robust processing tools** adhering to **international accredited standards** for the targeted languages, especially as regards tools that incorporate linguistic knowledge (e.g. morpho-syntactic taggers). It is noteworthy that even tools conformant to the same standard may exhibit differences: for instance, the MULTEXT-EAST (used by the SL and BG tagger) and the PAROLE tagsets (used by the EL tagger), both conformant to the EAGLES standard, were not entirely compatible with each other, making thus imperative the harmonization of the resources (through mapping procedures) to a common INTERA tagset. Naturally, conversion of output formats is preferable to the development of new tools, which is more time and cost demanding.

As a consequence, the scarcity of resources forced us to early abandon the ideal scenario of the production of a true multilingual parallel corpus, in favor of a ***comparable parallel corpus***, as described in section 2.1.1, a decision that had direct impact on the construction of the terminological resources.

3.2. Terminologies

3.2.1. State of the art

Semi-automatic procedures for terminology extraction usually consist in shallow techniques that range from stochastic methods to more sophisticated syntactic approaches (Jacquemin, 2001; Bourigault et al., 2001). All of them, however, converge in identifying terms mostly on statistical grounds, on the basis of their relative frequency in a corpus, possibly augmenting these measures with filters capturing the domain specificity of a term. Although not theoretically correct (as the status of "termhood" is in principle independent of the number of occurrences, and a hapax might well be a term), this practice is rooted in computerized terminology, where computer-aided text analysis and the possibility of processing large amount of information have changed the bases of terminology compilation, as well as the way in which term appropriateness is conceived and the degree of human intervention foreseen in this process. In this particular context, we adopted a *hybrid approach* to terminology extraction from multilingual parallel texts, *combining statistical and symbolic techniques*.

3.2.2. Implications of the corpus change for term extraction

The change in the final configuration of the INTERA multilingual parallel corpus as described above had obvious consequences to the task of term extraction.

The first consideration to be made concerns the *degree of multilingualism* of the terminology: since the multilingual corpus is not parallel across all languages, the final terminology is not a truly multilingual one. In other words, the lexicon is not the same across languages and terms are not all interconnected and corresponding to each other. Instead, for each EN–language*X* pair, we derived the corresponding terminology, thus arriving at a bilingual (EN-*X*) terminology for each domain. Since the domains are at least partially overlapping, some terms occurring in one terminology also occur in another one, thus enabling us to build truly multilingual links at least for a subset of terms, namely in the domain of Law.

The second consideration is related to the *range of technical solutions adopted for automatic term extraction*. The availability of the same pivot language for all target languages proved useful, especially because few reference corpora and NLP tools are available for the target languages. On the contrary, there is a huge amount of LRs (corpora, lexica and tools) available for the English language, and this allowed us to opt for a combination of statistical and NLP procedures, as illustrated in more detail in the next section.

[4] A similar position is hinted at in (Resnik & Smith, 2003) and (Mihalcea & Simard, 2005).

[5] The INTERA resources are to be distributed; therefore, IPR clearance has been a critical criterion during text selection.

3.2.3. Term extraction methodology taking into account reality

As introduced above, the data available for the task of automatic term extraction come under the form of four parallel corpora. Each parallel corpus is further organized according to the particular domain to which the texts of the corpora belong. The size of available data is important for determining the coverage of the terminological resource, since more data mean more terms. It is also important for the quality of the terminological resource, as the automatic procedure needs a statistically relevant amount of data to yield high-quality data. Unfortunately, the available data dramatically differed in size both across different domains and across different languages, thus yielding domain-specific terminologies very different in size and hence term coverage. This situation is clearly depicted by the case of the terminology for the health domain (see Table 2). The corpus data amount to 13 Mb for EL and 1Mb for SR. The larger quantity for EL allows to extract more candidate terms, as easily foreseen, but, most importantly, to produce less candidate translators and of better quality: while for EL the candidate translators/terms ratio is of 1,5, for SR it is of 4,1.

Extraction procedure

The task of automatic term extraction was organized in three main phases:

1. Automatic extraction of terms from the EN components of the parallel corpora;
2. Automatic identification of candidate translators in the target languages;
3. Manual validation of the candidate translators found with the automatic procedure.

Extraction of English candidate terms

Candidate single terms were extracted by comparing the relative frequency of lemmas inside each domain and language specific sub-corpus against a lemma-based frequency lexicon of the *British National Corpus*, which was used as a reference corpus. The comparison between the frequency distributions of terms in the general lexicon and that of the different domain-specific lexicons was performed adopting a mathematical function evaluating the distance of the frequency of domain-specific terms from the frequency which was expected on the basis of the general lexicon. We compared the lists generated adopting several different mathematical formulae, among which are the following:

$d1 = f_{r\ (specialized\ lexicon)} - f_{r\ (general\ lexicon)}$

$d2 = f_{r\ (specialized\ lexicon)} / f_{r\ (general\ lexicon)}$

$d3 = \log(f_{r\ (specialized\ lexicon)} / f_{r\ (general\ lexicon)})$, where f_r represents the relative frequency of a term inside the lexicon.

In order to select candidate multi-word terms, we specified a bunch of basic syntactic rules expressing constraints over syntactic patterns. In order to avoid over-generation problems, some corrective measures have been applied, most notably by specifying lists of words to be discarded *a priori* (stop-word lists) and by applying different values of the threshold under which a candidate is automatically discarded; the threshold is each time adjusted depending on the overall size of the parallel corpora under analysis and empirical measures.

Extraction of candidate translators

Once candidate terms were identified for EN, we turned to the task of automatic identification of candidate translators in the target languages (TL). To this end, we exploited the structural information available in the parallel corpora from which the terminology was to be extracted. Since the sentences in the TL texts were aligned to those of the pivot language (PL), it was easy to select a suitable search space for any candidate term. The algorithm for the extraction of candidate translators consists of the following steps:

a. Selection of the *source region set* from the PL corpus;
b. Extraction of *target region set* from the TL corpus;
c. Search Extraction of lemmas from target region set;
d. Ordering of the lemmas contained in the search target region set according to a *ranking function*;
e. Selection of candidates.

Given a candidate term *t* (in EN), the target region in the TL corpus was easily identified thanks to the aligned data: each region of the EN corpus containing term *t* was uniquely associated with a region of the TL corpus. Then, the lemmas from the target region set were extracted, filtering out lemmas of function words. It was observed that the TL lemmas could be classified on the basis of their "probability" of being a translation of a given term by means of simple frequency analyses. This classification is obtained through the synthesis of a ranking function. Several hypotheses were considered, all of them aiming at highlighting the statistical "idiosyncrasies" of the translating lemma. The best performing measure is the following:

$$f(l) = r(l) - q(l) * |I|$$

where $r(l)$ is the number of regions of the target region set containing at least one occurrence of lemma l, $q(l)$ is the ratio between the number of regions containing lemma l and the total number of regions in the corpus; $|I|$ is the total number of regions of the target region set.

Validation

The lists of candidate EN terms and their corresponding candidate translations in the other languages (lists of single word terms and multiword terms as produced by the tool) were presented to human validators, all native speakers of the selected languages. The final lists produced by the validators were used for the production of the multilingual terminological entries.

4. The production model

4.1. LRs production: the business

LRs production is an endeavour undertaken by a wide range of actors, these being academic units, research centres and institutes, or companies (software developers, translation agencies, localisers etc.) and for a wide range of purposes (development and/or evaluation of HLT applications, educational material, theoretical and applied research etc.). In most of these cases, LRs are designed and produced specific to the needs of the intended application or usage, either by following up-to-date and widely accepted standards or, often enough, in custom-made formats. An immediate consequence of this policy is the difficulty of portability to new applications/usages. Moreover, LRs-on-demand production is rarely accompanied by a specific plan for their marketing thereafter. Finally, the task is furthermore hampered by the diversity in the existing (or not) infrastructure (regarding raw data as well as processing tools), but also by cultural approaches related to the issue of language.

4.2. The LRs production model

The model proposed here concerns the production of MLRs and discusses the necessary steps to be undertaken in this process, in order to minimize effort and cost without jeopardizing quality.

4.2.1. Market-oriented development of LRs

The development of LRs should not ignore the market; this refers both to the existing situation and to potential users' needs. Although this approach seems incompatible with the infrastructural character of resource production, tailoring LRs production to the users' needs is considered crucial in order to overcome the strenuous and repetitive effort connected to this task. In other words, LRs production should cater both for specific application needs as well as for prospective (re-)uses of the same material.

A careful overview of the market would provide valuable insight as to (a) which types of LRs, domains, languages, and relevant metadata would meet prospective customers' needs, and (b) what LRs exist that could be re-used. The quality of the final resources produced is strictly intertwined with the existing source material, its (re)usability, the level of annotation of the material, or the availability of tools that perform adequate and robust processing, the compatibility/reusability of different encoding formats and annotation schemes etc.

The goals set based on this overview of the market, however, should be flexible, in order to accommodate possible shortcomings due to actual circumstances.

4.2.2. LRs collection issues

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

However, the re-usability approach is hampered by several factors. Indicatively, we could mention scarcity of related promotion activities and absence of efficient metadata descriptions and informative documentation, which render the existing resources difficult to locate. Moreover, the lack of uniformity in existing LRs encoding schemas combined with the use of a large variety of tools at different levels of processing leads to lack of interoperability and necessitates the extensive use or development of conversion tools.

As a source and, despite certain drawbacks (such as the one discussed above on the parallel nature of its material), the World Wide Web is indisputably a mine of language data of unprecedented richness and ease of access (Kilgarriff & Grefenstette, 2003).

The exploitation of advanced technologies (web crawlers, agents, language identifiers etc.) in the text identification and/or selection stage is imperative. Several techniques have already been employed to exploit the web content (Kilgariff & Grefenstette 2003; Resnik & Smith 2003; Mihalcea & Simard 2005); however, available tools and techniques may need fine-tuning and, sometimes, enhancement in order to meet specific needs of the intended application. For instance, available tools for the identification of potential parallel web texts work only on pairs of languages, which hinders the task of automatically identifying a parallel multilingual corpus. At the same time, the performance of these tools in discovering true parallel texts is not yet without problems; a suggested improvement would be the usage of alignment tools and techniques in the identification process, in order to check "parallelness" before downloading the texts. New tools should also be developed for the mining of documents uploaded to web sites, an important source of linguistic material yet unexploited.

Of course, the traditional method of exploiting raw and/or processed material supplied by data providers should not be overlooked: the web is not a panacea! A well-planned ahead strategy, based on the candidate provider's profile, should be carefully elaborated to ensure success of the endeavour: for instance, material from public organisations may present less IPR problems but demands more time in overcoming bureaucracy and spotting the right person(s) responsible for clearance of IPR and/or providing the textual material itself; on the other hand, private companies are usually less keen on providing their material without compensation.

IPR issues, already discussed in section 3.1.3, constitute a factor not to be taken lightly in the LRs production: one may easily download the entire web for their own personal research but cannot include a single text in an LRs package intended for distribution. Legal advice should be sought at the earliest project stages.

4.2.3. LRs processing issues

Serving as an illustration of the related issues and the process to be followed for LRs processing, we present here in more detail a model for parallel MLRs development. This is graphically presented in Figure 1, showing the technical stages of parallel MLRs production to be further used for multilingual terminology extraction, the tools needed and the types of data produced.

Parallel MLRs production can have two starting points: the utilization of raw data and the re-use of existing multilingual parallel data; each one with its advantages and disadvantages. Each starting point dictates a different pipeline of text-processing tools, which, however, converge on the output, namely, aligned parallel data, annotated at the structural and morphological level, suitable for term extraction. As obvious, re-use of existing data drastically reduces the text collection stage, but involves a (sometimes) heavy conversion stage, aiming to render the data conformant to the specifications of the task. On the contrary, starting from raw data necessitates all the stages of text processing, without, however, facing the problems of legacy data.

Important factors that contribute to the efficient interoperability and (re-)use of processed LRs include conformance to existing or emerging standards, as described in section 3.1.3, and clear distinction between the textual material per se and the annotation data; this latter becomes more important as we deal more and more with multi-level annotation, which may be further enriched at later stages, even after the completion of the initial LRs production project.

Finally, the use of available processing tools is considered preferable, for time and cost reasons, to the development of new ones, even if their output is not always conformant to the format and/or quality required. Again, the importance of adequate documentation and formal descriptions of such tools should be stressed here for their quick and efficient identification and use.

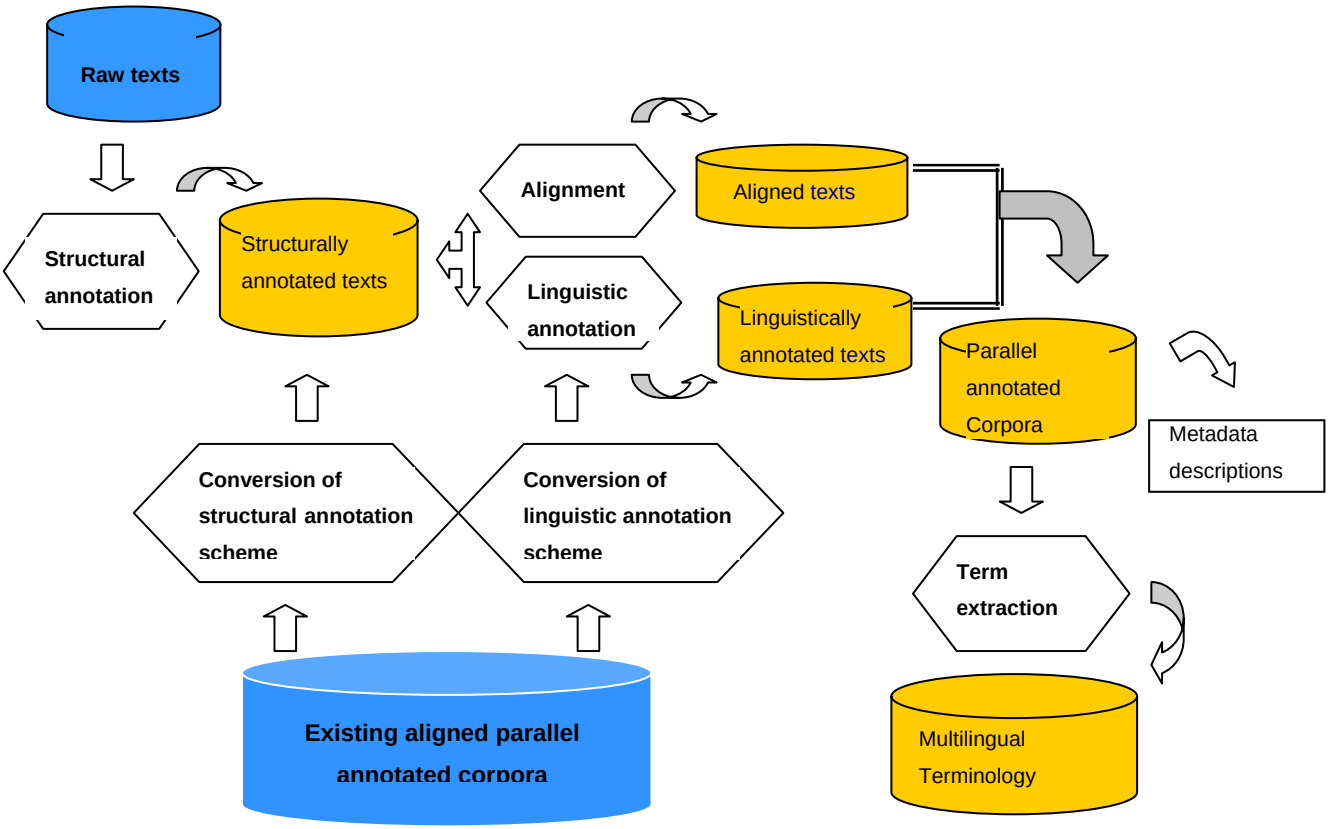

Figure 1 : Parallel MLRs processing model

4.3. Key points in the LRs production process

Independently of the application intended and the goals of the LRs production, key issues that allow for efficient development and re-use that should be taken into account when launching such an endeavor, are:

- adherence to existing or emerging standards in order to ensure interoperability,
- use of automatic or semi-automatic processes, updates and enrichment of resources,
- use of a distributive model of work,
- efficient metadata descriptions,
- dissemination and promotion activities,
- use of existing distribution channels.

5. References

Bourigault, D., Jacquemin, C., L'Homme, M.-C. (eds) (2001). *Recent Advances in Computational Terminology*. Amsterdam & Philadelphia: John Benjamins.

Calzolari, N., Choukri, K., Gavrilidou, M., Maegaard, B., Baroni, P., Fersoe, H., Lenci, A., Mapelli, V., Monachini, M., Piperidis, S. (2004). ENABLER Thematic Network of National Projects: Technical, Strategic and Political Issues of LRs. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*. Paris: ELRA

Desipri, E., Gavrilidou, M., Labropoulou, P. (2003). Language resources in the Balkan area. In *Workshop on Balkan Language Resources and Tools*, Thessaloniki, Greece.

Erjavec, T. (2004). MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*. Paris: ELRA.

Gavrilidou, M., Desipri, E., Labropoulou, P., Piperidis, S., Soria, C. (2003). *Technical specifications for the selection and encoding of multilingual resources*. INTERA Deliverable D5.1.

Gavrilidou, M., Desipri, E. (2003). Final Version of the Survey. ENABLER Deliverable 2.1.

Gavrilidou, M., Desipri, E., Labropoulou, P., Piperidis, S., Soria, C. (2005). Building multilingual terminological resources. In *Proceedings of the RANLP 2005 International Workshop on Language and Speech Infrastructure for Information Access in the Balkan Countries*. Borovets, Bulgaria.

Jacquemin, C. (2001). *Spotting and Discovering Terms through Natural Language Processing.* Cambridge, MA and London: The MIT Press.

Kilgarriff, A., Grefenstette, G. (2003). Introduction to the special issue on the web as corpus. *Computational linguistics*, 29(3), pp. 333 - 347.

LISA (2001). The LISA Globalization Strategies Awareness Survey. LISA.

LISA/AIIM (2001). The Black Hole in the Internet: LISA/AIIM Globalization Survey. LISA/AIIM.

LISA/OSCAR. 2003. Translation Memory Survey.

Maegaard, B., Choukri, K., Mapelli, V., Nikkhou, M., Povlsen, C. (2003). *Language resources - Industrial needs.* ENABLER Deliverable 4.2.

Mihalcea, R., Simard, M. (2005). Parallel Texts. *Journal of Natural Language Engineering*, 11(3), pp. 239-246.

Resnik, P., Smith, N. A. (2003). The Web as a parallel corpus. *Computational Linguistics*, 29(3), pp. 349 - 380.