The LRE Map. Harmonising Community Descriptions of Resources

Nicoletta Calzolari∞, Riccardo Del Gratta∞, Gil Francopoulo ‡, Joseph Mariani ‡, Francesco Rubino∞, Irene Russo∞, Claudia Soria∞

∞ILC/CNR
Pisa - Italy
‡ IMMI, LIMSI-CNRS
Paris - France
E-mail: {riccardo.delgratta, francesco.rubino, irene.russo, claudia.soria, nicoletta.calzolari}@ilc.cnr.it, {gil.francopoulo, Joseph.Mariani}@limsi.fr

Abstract

Accurate and reliable documentation of Language Resources is an undisputable need: documentation is the gateway to discovery of Language Resources, a necessary step towards promoting the data economy. Language resources that are not documented virtually do not exist: for this reason every initiative able to collect and harmonise metadata about resources represents a valuable opportunity for the NLP community. In this paper we describe the LRE Map, reporting statistics on resources associated with LREC2012 papers and providing comparisons with LREC2010 data. The LRE Map, jointly launched by FLaReNet and ELRA in conjunction with the LREC 2010 conference, is an instrument for enhancing availability of information about resources, either new or already existing ones, reinforcing and facilitating the use of standards in the community. The LRE Map web interface provides the possibility of searching according to a fixed set of metadata and to view the details of extracted resources. The LRE Map is continuing to collect bottom-up input about resources from authors of other conferences through standard submission process. This will help broadening the notion of "language resources" and attract to the field neighboring disciplines that so far have been only marginally involved by the standard notion of language resources.

Keywords: language resources, metadata, documentation

1. Introduction

Language Resources need accurate and reliable documentation because, if they are not documented, they virtually do not exist. However language resources are still often poorly documented or not documented at all. Use of metadata elements to describe and document resources is still uncommon and often inconsistent. It has been estimated that only 10% of existing resources are known, either through distribution catalogues or via direct publicity by providers (web sites and the like).

Single authors can find it difficult to document their own resources, simply because they can have a hard time deciding the relevant set of metadata elements to be used. Moreover, there is no sufficient awareness about the importance of documentation, that is often disregarded as a useless burden.

It is important, thus, to devise new ways for encouraging documentation of resources, and at the same time making it easy to perform.

The LRE Map of Language Resources and Tools is an initiative jointly launched by FLaReNet and ELRA in May 2010 with the purpose to develop an entirely new instrument for discovering, searching and documenting language resources, here intended in a broad sense as both data and tools. It was conceived as an instrument for capturing community knowledge about language resources, collecting descriptions both for tools and existing/ new resources applied in NLP research.

It was initially created in conjunction with the LREC 2010 Conference, as a campaign for gathering information about the language resources and technologies underlying the scientific works presented during the conference. Authors who submitted a paper were requested to provide information about the language resources and tools either developed or used; the initiative was successful, with more than 1990 resources descriptions. The required information was pretty simple and related to basic information about the type of the resource, the language and modality represented, the intended or real application purposes, the degree of availability for further use, the maturity status, the size, type of license and availability of documentation.

After LREC2010, thanks to massive input from the community, it has been possible to harmonise resources/ tools descriptions, finding out descriptive dimensions previously not available and now included in the metadata set that LREC2012 authors can use (see par. 3).

After this experiment, linking with other conferences (COLING, EMNLP, InterSpeech, Oriental COCOSDA, LTC, RANLP and others) was crucial to augment the information collected. By now the Map contains information about more than 3500 resources (data and tools) for 162 different languages and it complements existing cataloguing efforts (ELRA, LDC). Its main goal remains to gather information collected bottom up and to exploit the community knowledge helping the discovery and documentation of resources, essentially through a web interface (http://www.resourcebook.eu) that enables searches with multiple criteria.

During LREC2012 submission procedure 925 resources from main conference authors plus 276 from workshops authors have been collected and they are searchable in the LRE Map interface by selecting this conference in the conferences box.

2. The LRE Map interface

So far the LRE Map is available through a web interface designed to search and browse the data. The web interface currently provided to the community is a very simple interface based on normalised data and a login system to manage simple access management to resources. With respect to the previous release, it also offers a better visualization of data. The LRE Map web interface provides the possibility of searching according to a fixed set of metadata and to view/edit the details of extracted resources (edit if the user is directly related to the resource: i.e. when the user is an author of the paper that cites the resource). In addition, the database contains a lot of implicit relations, for example relations among authors (because in some way related to same resources) and resources (because cited by same authors in different papers).

3. Data Normalization

The normalization of new values performed on the basis of users' input makes the LRE Map a valuable source for the investigation of metadata, with the aim to clarify descriptive dimensions compatible with emerging trends in NLP.

During a first experimental phase a limited set of simple metadata helped collecting information in a fast and non-intrusive way during LREC2010 paper submission process. However it was expected that users' needs would go in different directions and the option "Other- specify" was made available for all the fields (Calzolari et al. 2010).

The set of metadata fields remains the same with respect to LREC2010 (see Table 1) because it was minimal but complete during that first experiment. However, for LREC2012 there is he novelty of size metadata splitted in two fields, one for numerical values and the other for the unit of measurements.

• Resource Type
• Resource Name
• Resource Production Status
• Use of the Resource
• Language(s)
• Modality
• Resource Availability
• Resource URL (if available)
• Resource Description
• Resource Size
• Resource License
• Resource Documentation

After LREC2010 it was clear that users input required normalization procedures. The aim was to reduce noise but also to highlight through manual inspection descriptive dimensions previously missing.

Different strategies have been employed for each metadata:

- **Resource Nam**e: the alternate forms are due to spelling or alternative names.
  When possible, each name has been normalised putting the available acronym in parentheses. Name variants have been preserved to help users during search through auto-completion (i.e. the name variant International *Corpus of Portuguese (CINTIL)* has been included together with the normalised version *Corpus Internacional do Português (CINTIL)*).
- **Resource Type:** among more than one hundred new values provided for this field, 6 have been chosen after discussion as new metadata in the list of suggested types ("Corpus Tool", "Language Resources/Technologies Infrastructure", "Machine Translation Tool", "Language Modeling Tool", "Spoken Dialogue Tool", "Text-to-Speech Synthesizer"), while 2 were merged ("Evaluation Methodology/ Formalism/ Guidelines" and "Evaluation Standard/Best Practice" are now "Evaluation Methodology/Standards/Guidelines") and one was renamed ("Transcriber" is now "Speech Recognizer/Transcriber").
  58 resource type values have been included in a list available through auto-completion (see Appendix A) when authors select the value Other-specify because they are relevant but less frequent ("Machine Learning Tool", "Stemmer", "Aligner", etc.)
- **Modality:** the value "Speech/Written" emerged as a relevant dimension previously unlisted and it is now included.
- **Resource Use:** starting with 533 uses provided by LREC2010, Coling2010 and EMNLP2010 authors, 258 new uses were considered useful. They were found through string similarity filters (included in Google Refine) and through manual inspection; we collected uses that were similar and excluded too specific uses. The vast majority (231) of new values for this metadata are now included in the auto-completion window that appears when selecting "Other - please specify". Among these values we list "Temporal Reasoning", "Grammar Engineering", "Sentence type identification" etc.
- **Language(s):** each resource language(s) value is now followed by the respective ISO-693-3 code from Ethnologue (www.ethnologue.com) to help users to unequivocally identify languages. Name variants for languages have been included.

4. LREC2012 Resources: descriptive statistics

As for LREC2010 data, we can analyze LREC2012 data according to different dimensions, focusing on single metadata or combining two or more metadata elements and looking for the correlations. However this year we can also provide a first temporal comparison looking at values for resources metadata provided by LREC2010

and LREC2012 authors.

4.1 Monodimensional analyses

Extracting information relative to single metadata helps to shed light on resources trends relative to types, uses, languages etc.

The Resource Type value refers to general descriptive categories and it enables reports on resources used or created. After data normalization on LREC2010 authors' input the list of resource type have been revised and updated with frequent values provided by authors. For this reason, in reporting the most frequent resources type for LREC2012 it is possible to only partially compare these data with LREC2010 data.

Type	N. instances	Trend 2012
Corpus	399	↔
Lexicon	108	↔
Annotation Tool	52	↑
Tagger/Parser	44	↓
Corpus Tool	39	n.a.LREC2010
Ontology	24	↔
Evaluation Data	23	↔
Machine Translation Tool	14	n.a.LREC2010
Terminology	11	↔
Grammar/Language Model	10	↔
Software Toolkit	10	?
Evaluation Tool	8	↔

Table 1: Most frequent types of resources

From Table 1 it is clear that corpora and lexicon remain the most frequently reported resources and that, globally, the distribution is the same for the two editions of LREC, with a slight difference for Annotation Tool and Tagger/parser and the insertion of new types missing for LREC2010 (Corpus Tool, Machine Translation Tool etc.)

Looking at the Resource Production Status (Table 2) the percentages of resources newly created is even higher with respect to LREC2010 (53% vs. 44%) while the percentage of existing resources is obviously lower (46% vs. 56%). If we found the field very active two years ago, today this trend is even more clear. Another difference is the proportion between finished and in progress resources among the newly created ones: if for LREC2010 only 32% were described as finished, for LREC2012 authors consider finished 45% of newly created resources.

Production Status	%
Existing-updated	9.8
Existing-used	36.5
Newly created-finished	24.2
Newly created-in progress	29.2
Not Applicable	0.3

Table 2: Resource production status for all resources

Concerning languages, we introduced the possibility to list more than one language (up to 6 fields), even if the vast majority of the described resources remains monolingual.

In fact 64.4% of the resources described are monolingual, 21.3% are bilingual, 7.9% are trilingual while there are several resources that list 4 languages (3.5%), 5 language (3%), and more than 5 (2.4%).With respect to LREC2010 there are less monolingual resources (they were 73% in the past edition).

In LREC2012 data there are resources in 85 different languages; in Table 3 a list of the 20 most cited languages is given.

Language	Citations
English	194
French	69
German	63
American English	43
Spanish	42
Italian	22
Japanese	21
Dutch	19
Polish	18
Portuguese	16
Swedish	16
Mandarin Chinese	14
Arabic	12
Czech	12
Croatian	11
Catalan	10
Bulgarian	9
Russian	9
Hindi	8
Standard Arabic	8

Table 3: The 20 most cited languages

In Table 4 percentages relative to modality values are reported. With respect to LREC2010 a mixed value has been added, i.e. Speech/Written. They are in line with LREC2010 data, with the exception of Multimodal/Multimedia that was 9%.

Modality	%
Written	79
Speech	5.1
Speech/Written	5.1
Not Applicable	4.4
Multimodal/Multimedia	4
Modality Independent	1.5
Sign Language	0.5
Other	0.4

Table 4: Modality values for all resources

Data about uses, that report on the main application/task for which a resource is used in the research paper, are quite varied, as for LREC2010. In the past edition 53% of the tags were provided by users. For LREC2012 we provided a list of "Other" values, manually selected from

users' past input, with the aim to help in the filling of the form and to promote harmonisation. This strategy is successful because for LREC2012 only 5.6% of input relative to uses have been inserted without choosing an existing tag, a significant improvement toward normalisation.
In Table 5 a list of the most frequent values is reported, with basic comparison with LREC2010 to highlight trends relative to uses.

Application	%	Trend 2012
Information Extraction, Information Retrieval	11.4	↔
Machine Translation, SpeechToSpeech Translation	11.1	↔
Language Modelling	6.8	↑
Acquisition	5.1	↑
Word Sense Disambiguation	4.4	↑
Document Classification, Text categorisation	3.7	↑
Named Entity Recognition	3	↓
Knowledge Discovery/Representation	2.8	↓
Dialogue	2.6	↓
Discourse	2.5	↓
Emotion Recognition/Generation	2.5	↓
Text Mining	2.2	↔
Speech Recognition/Understanding	1.7	↓
Semantic Web	1.4	↓
Dependency Parsing	1.1	n.a.LREC2010
Natural Language Generation	1.1	↔
Corpus Creation	1	n.a.LREC2010
Language Identification	1	↓

Table 5: Most frequent uses for all resources

Looking at resources availability, we provided several values that pertain to the different means by which resources are distributed. Table 6 shows as the vast majority are freely available, with a lower percentage with respect to LREC2010 (52% vs. 54%) and a lower percentage of resources available from the owner (22 % vs. 28%) while, for resources obtained from Data Center(s), the percentage is the same for the two editions of LREC.

Resource Availability	%
Freely Available	52
From Owner	22
From Data Center(s)	9.5
Not Available	4
Other	12

Table 6: Availability of all resources

Type of license, documentation and size constitute additional values that are optional and as a matter of fact they are less populated with respect to the other values. 44.4% of resources report information on documentation, while 79% report information on size and 48.8% report information on license.
Splitting entries on the basis of the Resource Production Status highlights values and features that characterize newly created resources. Among newly created resources prevail corpora and lexicon, as shown in Table 7 and the vast majority of them are freely available (see Table 8) even if the percentage is lower when compared with the overall set of resources (see Table 6).

Resource Type	%
Corpus	56.1
Lexicon	10.9
Annotation Tool	6.3
Evaluation Data	3.8
Corpus Tool	3.39
Ontology	2.9
Tagger/Parser	1.8
Grammar/Language Model	1.5
Evaluation Tool	0.9
Language Resources/Technologies Infrastructure	0.9
Machine Translation Tool	0.9
Terminology	0.9
Database	0.6

Table 7: Types of newly created resources

Resource Availability	%
Freely Available	47.19
From Owner	26.63
Not Available	7.24
From Data Center(s)	7
Other	11.94

Table 8: Availability of newly created resources

If we look at resource types for newly created resources, we can understand current trends in resources creation: in LREC2012 newly created resources we found more evaluation data and less tagger/parser. This trend is even clearer if we look at LREC2010 newly created resources, for which evaluation data was 1.6%. Instead, the distribution of existing resources types is very similar if we compare LREC2010 and LREC2012. Similarly, the availability of existing resources (Table 10) is not different with respect to the past edition of the same conference.

Resource Type	%
Corpus	36.9
Lexicon	15.6
Tagger/Parser	9.3
Annotation Tool	6.2
Corpus Tool	5.7
Machine Translation Tool	2.6
Ontology	2.6
Software Toolkit	2.08
Evaluation Data	1.8
Terminology	1.8
Language Modeling Tool	1.5

Named Entity Recognizer	1.3
Evaluation Tool	1.04
Language Resources/Technologies	1.04
Machine Learning Tool	1.04

Table 9: Types of existing resources

Resource Availability	%
Freely Available	64.1
From Owner	17
From Data Center(s)	11.2
Not Available	1.6
Other	6.4

Table 10: Availability of existing resources

5. Conclusions and Future Developments

The LRE Map holds an unprecedented potential for possible applications and uses. It is an instrument for enhancing availability of information about resources, either new or already existing ones through the LRE Map interface (par. 2). It is a measuring tool for monitoring various dimensions of resources across conferences (par. 4), thus helping to highlight emerging trends in language resource use and related language technology developments, by cataloguing not only language resources in a narrow sense (i.e. language data), but also tools, standards and annotation guidelines.

The potential of the LRE Map for becoming a powerful aggregator of information related to language resources was immediately clear, as was the possibility of deriving and discovering new combinations of information in entirely new ways. For example, the database underlying the LRE Map can yield interesting matrices of the language resources available for the various languages, modalities, or applications. Such matrices have been already used in META-NET and FlareNet to provide a picture of the situation of resources availability for the various European languages. (Mariani, J. & Francopoulo, G. 2011).

The LRE Map will be linked to the Language Library (Calzolari et al. 2011) through the description of resources and tools used and both will also be available through META-SHARE (www.meta-share.eu).

In the near future the LRE Map will continue collecting bottom-up input about resources from authors of NLP conferences. Providing information about resources could permanently become part of the standard submission process. This will help broadening the notion of "language resources" and also attract to the field neighbouring disciplines that so far have been only marginally involved in the description of used language resources.

The LRE Map wants to have an impact in reinforcing and facilitating the use of standards in the community by allowing registration of resources together with submission of papers for a conference, making most used/most adopted standards emerge. Finally, the LRE Map wants to promote active and personal engagement in documenting resources, encouraging a change in culture. It will pave the way to an entirely new tradition in the field of Language Resources and Technologies that ultimately may lead to the concept of publication and citation of language resources to give academic credit along the lines of publications of papers through normalisation of metadata values, towards consolidation of unique ways of referencing language resources and assessing their impact factor.

The development of an appropriate platform that enables harmonization and semantic interpretation of the acquired dynamic information, with a focus on the sustainability in provision of new language resources metadata, is among the long term objectives.

6. Acknowledgements

We want to thank all the LRE Map contributors that provided accurate descriptions of existing and newly created resources and tools. Without their contribution the picture of resources usage would be poorest.

We thank the META-NET project (FP7-ICT-4 249119: T4ME-NET) for supporting this work. The LRE Map started as an initiative within FLaReNet - Fostering Language Resources Network.

7. References

Calzolari, N., Soria, C., Del Gratta, R., Goggi, S., Quochi, V., Russo, I., Choukri, K., Mariani, J., Piperidis, P. (2010), "The LREC Map of Language Resources and Technologies". In Proceedings of LREC 2010, pp. 949-956.

Calzolari, N., Del Gratta, R., Frontini, F., Russo, I. (2011), The Language Library: Many Layers, More Knowledge. Proceedings of Workshop on Language Resources, Technology and Services in the Sharing Paradigm, pages 93–97, Chiang Mai, Thailand, November 12, 2011, 93-97.

Mariani, J. and Francopoulo, G. (2011), D11.1.1 First Public Version of the META-Matrix available at http://www.flarenet.eu/?q=META-Matrixes.

Appendix A- New resource types listed as Other and available through auto-completion

Controlled Legal Language
Repository of bilingual lexicons
Resources integration
Query language
Geoparsing engine
Adaptation system
Application for Semantic Desktop
Cultural Graph Comparator
Data Entry System
Digital library management system
Lexical Isolation Point predictor
Parallel grid execution environment for HLT tools
Sentence Splitter
Spelling Corrector
Text Mining System

Text Navigation Tool
Text simplification tool
Tool for mapping language resources and users
Tool for transcribing scanned text
UIMA Toolkit
Dictionary
Thesaurus
Course material
3D toolkit
Acquisition Tool
Aligner
Chunker
Concordancer
Coreference Resolution
Corpus Tool
Game
Handwritten/Character Recognition Tools
Information Retrieval Tool
Knowledge Processing Tool
Language Modeling Tool
Language Processing Infrastructure
Lemmatizer
Lexicon Tool
Machine Learning Tool
Machine Translation Tool
Morphological Analyzer/Generator
Ontology Tool
Question Answering Tool
Search Engine
Sentiment Analysis Tool
Software Toolkit
Segmentation Tool
Speech Recognizer
Spoken Dialogue Tool
Stemmer
Summarizer
Talking Head
Terminology Tool
Text-to-Speech Synthesizer
Textual Entailment Tool
Transliterator
Web Service
Wikipedia Tool