ECP-2007-LANG-617001

FLaReNet

The FLaReNet Databook

Deliverable number/name	*The FLaReNet Databook*
Dissemination level	*Public*
Delivery date	*22 October 2011*
Status	*Final*
Editors	*Paola Baroni, Claudia Soria, Nicoletta Calzolari*
Contributors	*Victoria Arranz, Núria Bel, Gerhard Budin, Tommaso Caselli, Khalid Choukri, Riccardo Del Gratta, Elina Desypri, Gil Francopoulo, Francesca Frontini, Sara Goggi, Olivier Hamon, Erhard Hinrichs, Penny Labropoulou, Lothar Lemnizer, Steven Krauwer, Valerie Mapelli, Joseph Mariani, Monica Monachini, Jan Odijk, Jungyeul Park, Stelios Piperidis, Adam Przepiorkowski, Valeria Quochi, Eva Revilla, Laurent Romary, Francesco Rubino, Irene Russo, Helmut Schmidt, Hans Uszkoreit, Peter Wittenburg*

e*Content*plus

This project is funded under the *e*Content*plus* programme[1],
a multiannual Community programme to make digital content in Europe more accessible, usable and exploitable.

[1] OJ L 79, 24.3.2005, p. 1.

1. Introduction

Recognizing the lack of information about existing language resources as one of the major factors hindering the development of the field, FLaReNet has undertaken a number of actions to survey existing resources, inform about them, and enhance their visibility. The fieldwork carried out by the project has gathered and revealed a set of facts and figures concerning the field of Language Resources and Technologies (LRT) (existing language resources, their production models, current and existing activities in the various countries, the gaps with respect to resources and technologies, a panorama of existing standards and plans for their adoption, etc.). FLaReNet, in compliance with its commitment to provide a picture of the current state of the LRT field, believes that making all quantitative data available public and open to the community is of utmost importance.

The FLaReNet Databook is not only the collection of all the factual material collected during the activities of the project, but also a set on innovative initiatives and instruments that will remain in place for the continuous collection of such "facts".

The purpose of the Databook is in fact, on one side, to consolidate the analyses carried out in the project and, at the same time, to set up the proper mechanisms that will enable the provision of a continuous stream of relevant factual material, also after the end of the project.

As such, it is intended both to give a snapshot of the *status quo* at the time of the issue of the final FLaReNet Blueprint of Actions and Infrastructures[2], and – most importantly – to establish new methods and new systems to carry on identification of facts and information, and even collaborative creation or enrichment of new language resources.

Data varies considerably in type, internal organization, intended addressees and fruition mode. A major distinction can be drawn between community-oriented data – that will have a life beyond the project – and project-oriented data.

Data and information designed and built for the community are immediately usable by external users, even if with different degrees of user-friendliness, since they range from dedicated interfaces (§ 2.1) to structured documents (§§ 2.2 and 2.4) up to collaborative or wiki-like repositories (§§ 2.1, 2.3, 2.5 and 2.6). Some of these data is already extensively used and has become an important asset for all the community; other has just started and may be the first step towards innovative ways of collaborative resource creation.

Project-oriented data (§§ 2.7 and 2.8), was used for project internal purposes only – to prepare the initial surveys – and is relatively unstructured and in raw format.

The Databook is made publicly available by FLaReNet in digital format at the following URL: http://www.flarenet.eu/?q=FLaReNet_Databook.

[2] See chapter 1 of the final FLaReNet Book (http://www.flarenet.eu/sites/default/files/FLaReNet_Book.pdf).

2. Content and Structure of the FLaReNet Databook

The following sections provide a concise description of the various types of data that together constitute the FLaReNet Databook. Some of this data is also described in a number of publications.

As said above, the data is distinguished in two major types of data: community-oriented and project-oriented.

From each section-title there is a link to the data themselves.

Community-oriented Data

2.1 LRE Map of Language Resources and Tools

The LRE Map is an instrument for capturing the community knowledge about Language Resources and Technologies. Its goal is to exploit the collective knowledge to help the discovery and documentation of resources, by promoting active and personal engagement in documenting resources, and thus also encouraging a change in culture.

The LRE Map was introduced as a completely new instrument at LREC 2010, and since then participants in various LRT events were asked to contribute a description of the resources, either used or produced for their research, presented at conferences, such as LREC, COLING, EMNLP, Interspeech, RANLP, EAMT, IJCNLP, ACL-HLT and Oriental COCOSDA. A limited set of simple metadata helped to collect reliable and accurate information in a fast and non intrusive way.

The Map contains information about more than 2700 resources (data and tools) for 162 different languages: their type, modality, availability, intended or actual application purposes, innovation potential, stage of development, etc.

The LRE Map holds an unprecedented potential for possible applications and uses. As an instrument for enhancing the availability of information about resources, either new or already existing ones, in conjunction with other catalogues already available from Data Centres. As a measuring tool for monitoring various dimensions (metadata elements) of resources across places and times, the LRE Map will help to highlight evolutionary trends in the language resource use and the related language technology development. By cataloguing not only language resources in a narrow sense (i.e. language data), but also tools, standards, and annotation guidelines, the LRE Map will help to broaden the notion of "language resources", and thus attract towards the field neighbouring disciplines that so far have been only marginally involved by the standard notion of language resources. By making most used/adopted standards emerge, the Map will have an impact in reinforcing and facilitating the use of standards in the community.

By allowing the registration of resources together with the submission of papers for a conference, it will pave the way for an entirely new tradition in the field of LRT, which ultimately may lead to the concept of publication and citation of language resources to give academic credit along the lines of publications of papers. Finally, through the normalisation of metadata values, the Map will pave the way for the consolidation of unique ways of referencing language resources.

2.2 META-Matrixes

The objectives of the Language Matrixes (built with the data of the LRE Map also within META-NET, thus showing the breeding of a project into another) are to provide a clear picture of what exists in terms of language resources, in the broad sense (data, tools, evaluation and meta-resources) for the various languages, and to stress the languages that are missing such language resources. The goal would then be to ensure the production of the corresponding resources to fill the gaps for those languages.

The Language Matrixes provide an easy way to get that picture and to have access to the details of the corresponding resources.

The Language Matrixes were built automatically on the basis of the LREC Map data, and therefore from the information provided by the authors of the papers submitted at LREC 2010, which gathers the international community working in the field of LRE.

In this first analysis, we considered the 23 official languages of the European Union, together with a category on "Regional European languages" and one on "Non-EU European languages", as well as "Multilingual", "Language Independent" and "Not Applicable" categories. We produced 8 Language Matrixes on: Multimodal/Multimedia Data and Tools, Written Language Data and Tools, Spoken Language Data and Tools, Evaluation and Meta-resources (standards, metadata, guidelines). Several types of resources are listed for each matrix, corresponding either to the 24 types that were suggested in the questionnaire or to the author's own entry, when no suggested type was found appropriate. The outcome is a total of about 150 types of resources, with a variable number for each matrix (from 5 Types for Evaluation to 78 Types for Written Language Tools).

Those matrixes show that the English language is by far the most resourced language (not surprising), followed by French and German, Spanish, Italian and Dutch. Some languages are clearly under-resourced, such as Irish Gaelic, Slovak or Maltese. Given the large number of Types expressed by the conference authors, some of them may exist only for one language, and the matrixes therefore show a large number of zeroes for all the other languages. We however preferred to keep that information as such rather than merging them into an "Other Type" category, as those singletons may be weak signals announcing a new research trend. Another option is to merge those singletons into a single "Other" category to facilitate the browsing of the Language Matrixes.

Since the LREC Map produced at LREC 2010, more data have been harvested at other conferences, which will be included in the next versions of the Language Matrixes, as well as the language resources appearing in journals, such as the *Language Resources and Evaluation* journal, and in LR catalogues, such as the LDC or ELRA ones. Building the Language Matrixes from actual data provided by authors allows over time to reflect a landscape that is continuously evolving with more and more language resources. Within META-NET, the Language Matrixes have already started being used for identifying the Language Gaps and for writing the Language Tables in the Language Reports.

2.3 Feedback from Contact Points on National Initiatives in the Area of Language Resources

In order to conduct a survey of the National and Transnational initiatives in the area of LRs, a network of international FLaReNet Contact Points was created in August 2010, comprising 102 colleagues from

78 countries or regions that are subdivided as follows: 26 EU member countries, 6 EU regions, 9 non-EU European countries and 37 non-European countries.

The survey shows that almost all European countries now take care of gathering language resources for their languages in order to conduct research investigations and develop and test systems for those languages. The languages which were considered as "low-resourced" are starting to recover even if they still need many more language resources, given that no language has enough language resources available for the needs of the research and industrial communities. Surprisingly, UK has no National program for (British) English. The reason may be the importance of US activities regarding the processing of the (American) English language. Baltic and Nordic countries are conscious of the importance of language resources for the promotion and survival of their languages, and they accordingly have a policy to support that area, including for minority languages. There is also an important activity in some EU regions, either for specific languages (Basque, Catalan), or in general (the Trento region in Italy).

Activities in the other parts of the world are also impressive. Governmental initiatives in India or South Africa to cover the development of language technologies for all the official languages of those multilingual countries in order to meet the needs of all citizens is exemplary. The creation of Associations for the specific development of language technologies for the Arabic language or for African Languages is also an interesting trend, while Asia keeps on organizing the activities in the various countries, with individual initiatives dealing with the cultural heritage, the preservation of the languages spoken in the country being part of it.

2.4 FLaReNet Standards' Landscape Towards an Interoperability Framework

This document proposes an overview of the current scene towards an Interoperability Framework and acts as a reference point for the current standards that the community fosters and encourages to adopt/improve. This initiative is in close synchronization with other relevant initiatives such as CLARIN, ELRA, ISO, TEI and META-SHARE.

The document builds on the CLARIN Standardisation Action Plan and adapts and extends it to the needs of the broader LT Community, beyond the SSH research areas including the industry.

The main goal of this document is to give a practical orientation for various LT players, both commercial and academic; the main message being that a harmonized domain of language resources and technology can be achieved stepwise, but that an effort to adopt standards is necessary to overcome fragmentation.

This is to be intended by no means as a static, closed document, but rather a dynamic one which needs to be constantly/periodically revised and updated by the community itself. This document will also be an essential reference document for the META-SHARE initiative.

2.5 FLaReNet Repository of Standards, Best Practices and Documentation

FLaReNet has promoted the creation of a shared repository with standards, best practices, data formats, annotation schemes/tools and documentation of most well-known language resources. These materials, easily accessible by everyone, are seen as a first step to overcome current problems in the

production of language resources, thus enhancing efficiency, quality and interoperability, and promoting standardisation.

The current repository is just a first prototype, but a future establishment of an extensive (virtual) repository is seen as an important concrete step in promoting standardisation and adoption of best-practices, in view of the needed interoperability among language resources and language technology.

2.6 LREC Language Library[3]

After the success of the LRE Map introduced in LREC 2010 – now used in many conferences as a normal step in the submission procedure – FLaReNet and ELRA have launched for LREC 2012 the LREC Language Library. The LREC Language Library is intended to be a collaborative enterprise of the LRT Community, an important contribution to a "community-built" Open Resource Infrastructure.

An LREC Repository has been prepared, hosting a number of resources on all modalities (speech, text, images, etc.) in as many languages as possible. When submitting a paper, authors are invited to process some pieces of this data, in the language(s) of choice, in one or more of the possible dimensions addressed by the submission (e.g. POS-tag the data, extract/annotate named entities, annotate temporal information, disambiguate word senses, transcribe audio, etc.).

The processed data are then put back in the Repository, that will become one of the META-SHARE repositories, and will be made available to all the LREC participants before the conference, to be compared and analysed.

This collaborative work on annotation/transcription/extraction/… over the same data (many parallel or comparable) and on a large number of processing dimensions will set the ground for a large Language Library, linked to the LRE Map for the description of the data, where everyone can deposit/create processed data of any sort – all our "knowledge" about language.

The LREC Language Library will be made publicly available as soon as it is ready.

Project-oriented Data

2.7 FLaReNet WP3 Contribution

The data presented in this section have been collected as part of the work conducted for the The FLaReNet deliverable *D3.1 – Report on the scientific, organizational and economic methods and models for building and maintaining LRs*.

More specifically, the survey conducted in the framework of WP3 aimed at:

- creating a considerable and representative catalogue of existing Language Resources that are closely related to the HLT (Human Language Technology) field,
- identifying a wide range of their descriptive characteristics, development methodologies and management practices,

[3] The instructions for contributing to the LREC 2012 Language Library can be found at the following URL: http://www.lrec-conf.org/lrec2012/?LREC-2012-Language-Library.

- organizing the information in a searchable form and ensuring that the descriptive features (metadata) used for the LRs are reusable.

2.8 FLaReNet WP6 Contribution

The FLaReNet deliverable *D6.1a – Survey and assessment of methods for the automatic construction of language resources. Report on automatic acquisition, repurposing and innovative proposals for collaborative building of LRs* aimed to provide an overview of the state of the art of language resources in Europe, focusing in particular on their automatic construction.

The first part of the deliverable presented a survey of the most demanded resources, as core elements of many NLP applications, whereas the second part was devoted to an overview and quantitative analysis of the current techniques for automatic construction of language resources, including a section on last academic proposals for automatic acquisition and production of language resources in order to identify the most recent trends in this field, where the research community concentrates more efforts, and monitor the existence and status of methods or tools for the automatic acquisition of language resources.

To carry out this survey of last academic proposals we reviewed the research papers presented at some of the major international conferences for computational linguistics and language resources from 2006 to 2009 (ACL, COLING, LREC and EACL).

This contribution presents a classification of the scientific papers reviewed for the production of D6.1a. The classification is based on the types of resources to be produced or the type of linguistic information to be acquired.

2.9 References

N. Calzolari, C. Soria, R. Del Gratta, S. Goggi, V. Quochi, I. Russo, K. Choukri, J. Mariani, S. Piperidis, 2010. "The LREC Map of Language Resources and Technologies", in N. Calzolari, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, D. Tapias (Eds.), *Proceedings of LREC 2010 - 7th International Conference on Language Resources and Evaluation (Valletta, Malta, 17-23 May 2010)*, European Language Resources Association (ELRA), pp. 949-956

N. Calzolari, C. Soria, R. Del Gratta, 2010. "The LREC 2010 Map of Language Resources and Tools", CLARIN Newsletter No. 9-10, pp. 10-11

J. Mariani, 2011. Introduction of the "Identification and Tracking of Language Resources" session, in Proceedings of the 3rd FLaReNet Forum "Language Resources in the Sharing Age - the Strategic Agenda", Venezia, Italy, 26-27 May 2011

J. Mariani, 2011, "A journey from LRE Map to Language Matrixes", in Proceedings of the 3rd FLaReNet Forum "Language Resources in the Sharing Age - the Strategic Agenda", Venezia, Italy, 26-27 May 2011

J. Mariani, C. Soria, 2011, "Identifying and networking forces: an international panorama", in Proceedings of the 3rd FLaReNet Forum "Language Resources in the Sharing Age - the Strategic Agenda", Venezia, Italy, 26-27 May 2011

N. Calzolari, "Opening the Language Library: let's build it together!". Presentation held at the third FLaReNet Forum "Language Resources in the Sharing Age - the Strategic Agenda", Venezia, Italy, 26 May 2011

J. Mariani, C. Soria, P. Baroni, N. Calzolari, Survey of the National and Transnational Initiatives in the Area of Language Resources, FLaReNet Wiki, June 2011

N. Calzolari, C. Soria (eds.) Submitted. “Language Resources of the future – The future of Language Resources. The Strategic Research Agenda”. Cambridge Studies in Natural Language Processing, Cambridge University Press

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