# A Guide for the Production of Reusable Language Resources

**Victoria Arranz, Franck Gandcher, Valérie Mapelli and Khalid Choukri**
ELDA (Evaluations and Language resources Distribution Agency)
55-57 rue Brillat Savarin, 75013 Paris, France
E-mail: arranz@elda.org, gandcher@elda.org, mapelli@elda.org, choukri@elda.org

### Abstract

The project described in this paper is funded by the French Ministry of Research. It aims at providing producers of Language Resources, and HLT players in general, with a guide which offers technical, legal and strategic recommendations/guidelines for the reuse of their Language Resources. The guide is dedicated in particular to academic laboratories who produce Language Resources and may benefit from further advice to start development, but also to any HLT player who wishes to follow the best practices in this field. The guidelines focus on different steps of a Language Resource "life", i.e. specifications, production, validation, distribution, and maintenance. This paper gives a brief overview of the guide, and describes a) technical formats, standards and best practices which correspond to the current state of the art, for different types of resources, whether written or spoken, at different steps of the production line, b) legal issues and models/templates which can be used for the dissemination of Language Resources as widely as possible, c) strategic issues, by offering a dissemination plan which takes into account all types of constraints faced by HLT community players.

## 1. Introduction

### 1.1 Context

As information search systems evolve, the Language Resources needed become more and more sophisticated, thus requiring considerable development efforts. As a matter of fact, multilingual Language Resources are required nowadays, given that neither English nor local languages are sufficient to enable an efficient internet watch in the field of scientific and technical information.

High quality Language Resources, whether for written or spoken language, are being built by academic research centres, as well as by industrial organisations. Nevertheless, a number of obstacles need to be overcome in order to enable their distribution and reuse by any third party. It is necessary to prevent such obstacles as well as to take into consideration every step in the life of a resource (production, identification, distribution, etc.) in order to make it available afterwards.

The obstacles to be highlighted are of a threefold nature:

- Technical:
  - The existence of multiple encoding formats and data storage conventions (e.g. SAM, Sphere, WAV, etc.) as well as conversion tools (UTF-8, ASCII, etc.),
  - The use of resource description formats (metadata), which are incompatible or do not comply with the state of the art (non standard proprietary formats). Such formats are needed to make inventories or to search for resources available at different places (ELRA Catalogue, LDC Catalogue, IMDI, OLAC, etc.),
  - The use of resource formats which are incompatible or do not comply with the state of the art (non standard proprietary formats).
- Legal:
  - The lack of legal concern within academic centres, who "omit/forget" to ask for prior authorisations,
  - The use of legal models with unduly restrictive distribution rights,
  - The different strata of intellectual property rights, which are not taken into account (e.g. the production of new resources which integrate resources already covered by intellectual property right),
  - The multiple home-made license models, often inspired by software license models (such as GNU, GPL, Creative Commons), in general not adapted to Language Resources,
  - The diversity of legal protection modes in Europe and over the world.
- Strategic:
  - The cost of adaptation or acquisition, which may not be in agreement with a certain « market » or with the financial capacity of its potential users,
  - The absence of a specific resource for a given need,
  - The unavailability of existing resources that the owner/provider may not want to share/distribute for market competition reasons (whether technological, strategic or financial).

ELDA (Evaluations and Language resources Distribution Agency) has been dealing with this type of issues since its creation in 1995, as the operational body of ELRA (European Language Resources Association). Those crucial issues are part of our original missions: Language Resource identification, collection, production, validation and distribution. For instance, with respect to legal issues, ELDA has worked with several lawyers in order to define template licenses dedicated to Language Resources. In this context, different types of licenses were drafted to handle, on the one hand, the relationship between data owners and distributor, and, on the other, the relationship between distributor and data users. This work enabled

ELDA to consolidate and control the exchange and sharing of data in a stable manner. This was materialized through the signature of several thousands of licenses between ELDA and data owners/users.

## 1.2 Objectives

The project described in this paper is being financed by the French Ministry of Research, and it aims at providing producers of Language Resources, and all HLT players in general, with a guide which offers technical, legal and strategic recommendations/guidelines for the reuse of their Language Resources.

For all types of resources, whether written or spoken, this guide aims to describe:

a) Technical formats, standards and best practices which correspond to the current state of the art.
b) Legal models/templates which can be used for the dissemination of Language Resources in as wide a manner as possible.
c) On a strategic point of view, this document aims to encourage collaboration and sharing of Language Resources, by offering a dissemination plan which takes into account all types of constraints faced by HLT community players.

The guide is dedicated in particular to academic laboratories who produce Language Resources and may benefit from further advice to start development, but also to any HLT player who wishes to follow the best practices in this field. The guidelines focus on different steps of a Language Resource “life”, i.e. specifications, production, validation, distribution, maintenance, and all of them with regard to the three major dimensions mentioned above: technical, legal, and strategic aspects.

In order to accomplish these objectives, three main tasks have been defined within the project. A primary task consists in identifying and compiling all different options available on the market with regard to technical, legal or strategic issues. A second task focuses on the actual production of a detailed guide, whose aim is to serve as a repository of recommendations/best practices that must be taken into account for the production and distribution of Language Resources. Finally, this guide is meant to be disseminated at large to all potential users, among others, by means of the Technolangue.net portal, already maintained by ELDA.

The findings and currently on-going work for the development of this guide of production are further detailed in the coming sections (Gandcher et al. 2008).

# 2. Technical issues

## 2.1 Standards for the production of Language Resources

When we come to the production of Language Resources, we can find a large number of guides and standards already implemented and used for different types of Language Resources, and at all levels of the production line. Those guidelines have enabled us to draw up a list with the different steps to be followed for the production of Language Resources. Once a production scheme has been defined (*specifications*), it is then required to *encode* and *store* the data in standard formats which will allow their reuse and their connection in other systems or applications. Beyond file formats, we also elaborate on best practices for data sharing and exchange, such as the use of different character sets. Finally, a now inevitable issue stands in the *validation* of LRs, based on a number of agreed upon criteria.

In this paper, we present some existing guidelines for four main types of LRs: Written Corpora, Written Lexica, Speech Resources and Multimodal/Multimedia Resources.

### 2.1.1. Specifications

For written LRs (corpora and lexica), as well as for Speech LRs the main standards which originally enabled to draw up specification criteria for the production of LRs were mostly defined within the EAGLES[1] project (the Expert Advisory Group on Language Engineering Standards, 1993). Other projects, in particular PAROLE-SIMPLE [2] and MULTEXT for lexica and corpora, enabled to update and adapt those specifications for more specific purposes. More recently, a large consortium, ISO TC37/SC4[3], developed working groups to reflect on various topics around Written LRs, in particular WG3 Multilingual Text Representation and WG4 Lexical Database.

As far as Speech LRs are concerned, several projects could implement the EAGLES guidelines. Before EAGLES, we can quote several guidelines resulting from big international projects such as the SAM report, funded by the European Commission since 1987 in the framework of the ESPRIT programme. This report was followed by the SpeechDat projects (or « family »[4]), also funded by the European Commission (SpeechDat(M), SpeechDat(II), Speecon, SALA I et II, Orientel, etc.).

As for multimodal resources, several projects contributed to the definition of best practices for the collection of multimodal corpora. Most of those collections were mainly dedicated to audio and video recordings in closed meeting rooms (seminars, lectures, interactive meetings). The first corpora focused on audio material, such as ICSI (Janin & Baron, 2003) and ISL (Burger et al., 2002) corpora. International projects then introduced multiple works on audio and video channels, such as CHIL (Computers in the Human Interaction Loop)[5], VACE (Video Analysis Content Extraction)[6], AMI (Augmented Multiparty Interaction)[7] or NIST Smart Space[8].

[1] http://www.ilc.cnr.it/EAGLES/home.html
[2] http://www.elda.org/catalogue/en/text/doc/parole.html
[3] http://www.tc37sc4.org
[4] http://www.speechdat.org
[5] http://chil.server.de
[6] https://control.nist.gov/dto/twiki/bin/view/Main/WebHome
[7] http://www.amiproject.org
[8] http://nist.gov/smartspace

### 2.1.2. Encoding and Storage

Again, we can gather some encoding and storage guidelines that have been created both for Written Corpora and Lexica. First, the ASCII code is a prevailing format for un-annotated data. Then, when needing to add linguistic annotation, SGML or XML are considered as most commonly used. For written data, the main encoding guidelines were built within the TEI (Text Encoding Initiative)[9]. Within the MULTEXT project[10], associated to EAGLES and with the collaboration of VASSAR/CNRS[11], an encoding standard for corpora was built, the CES (Corpus Encoding Standard), in order to define minimal encoding conventions, based on the TEI. As for character encoding formats, UTF-8 is the most commonly used. When dealing with speech resources, we can refer to file formats such as WAV, SAM and Sphere for audio files and ASCII for transcription files. For video recordings, the CHIL project used sequences of compressed images in JPEG format. However, MPEG video format is also widely used. For instance, TRECVID (TREC Video Retrieval Evaluation) [12] evaluation campaigns used MPEG-1, or NIST used MPEG-2 or its proprietary format NIST-SMD. Within the AMI project, we can also find video data with DIVX AVI format.

### 2.1.3. Validation

For all types of Language Resources, validation work is based on three main criteria: documentation, formal validation and linguistic content validation. A number of those validation formalisms have been gathered, developed or extended through the ELRA Validation Committee (VCom). Public web pages on the ELRA[13] web site are dedicated to such standards. There, validation manuals are provided for Written Corpora (McEnery et al. 1998), Written Lexica (Fersøe 2004) and Speech Resources (Van den Heuvel et al. 2000). Most of the validations are done manually, in order to avoid any mistakes. For Multimodal resources, the validation issue appeared more recently. Within the CHIL project, an internal procedure was defined for raw data based along the following validation lines (Moreau et al., 2007b): video data collection, "microphone arrays" audio data, other microphone audio data. For annotated data, another internal procedure was also implemented, mainly considering audio and video annotations distinctly, as well as checking all required documentation (Moreau et al., 2007a).

## 2.2 Standards for the dissemination of Language Resources

A number of different Language Resource description forms are currently in use on the market and have been studied. Those may be used, for instance, for the cataloguing of searching of resources available at different places. Among the existing formats, it is worth mentioning the following:

- OLAC (*Open Language Archives Community*)[14]: a community for the creation of a virtual library of international Language Resources, with very rich discussions on metadata development and improvement.
- IMDI (*International Standards for Language Engineering Metadata Initiative*)[15]: an initiative for the standardisation of metadata used to describe Language Resources,
- ELRA Catalogue of Language Resources [16] : a catalogue which gathers over 900 Language Resources described in a formalised way, implemented by the European Language Resources Association (ELRA),
- LDC Catalogue [17] : a catalogue of Language Resources mainly produced by the LDC (Linguistic Data Consortium), some of which coming from projects funded by the USA government.

Those different formats are detailed in the guide, as well as the most exhaustive list of any other available formats. The guide details both the convergence and divergence points among the formats. It also gives their classification according to their application field. To sum up, we present a number of format recommendations in order to ensure the most compatible, compliant and reusable descriptions.

# 3. Legal issues

When considering a Language Resource, we distinguish the following types of players: rights owner(s) –of one portion of the resource (e.g. a speaker who gave his/her voice in the case of speech resources) or of the whole resource, providers (which may be the owner itself), distributors, integrating developers and end-users. With the aim of making a resource available, one needs to define the relationship between those different players, through adequate legal agreements, at the proper time.

As a first step, our guide also takes into account legal and ethical issues related to intellectual property rights.

Moreover, we have drawn up an inventory of different licenses which are being used in the field. The fact that most existing licenses are « home-made », and most of the time inspired from software distribution models (such as GNU, GPL, Creative Commons) should be emphasised. Therefore, our guide includes an analysis of their compliance with the reuse of Language Resources.

Based on this information, and together with ELDA's experience, we have defined a number of criteria to propose a set of licensing recommendations, and have drafted some templates of licenses that can be used between the different players of a Language Resource production/distribution process.

[9] http://www.tei-c.org/index.xml

[10] http://aune.lpl.univ-aix.fr/projects/multext

[11] http://www.cs.vassar.edu/~ide/research/

[12] http://www.itl.nist.gov/iaui/894.02/projects/trecvid/

[13] http://www.elra.info

[14] http://www.language-archives.org

[15] http://www.mpi.nl/IMDI

[16] http://catalog.elra.info

[17] http://www.ldc.upenn.edu/Catalog

### 3.1 Relevant rights in the production and dissemination of Language Resources

An important prerequisite is to provide an accurate legal definition of what “language resources” may consist in; this definition is to determine the identification of rights and obligations with respect to the considered 'language resources'. On this account, we offer a classification based on a two-axis distinction: the nature, from a legal perspective, of the primary resource and the eventual resulting legal protection ; the existence and degree of integration at work and the corresponding possibility to differentiate multiple layers within the resulting resource. We will therefore mainly distinguish:

- raw un-annotated corpora making for the usual primary language resource,
- derivative language resources characterized by the adding of variable amount of linguistic comments,
- and what is strictly defined as a ‘database’ under French Law.

The existence, under French Law, of a specific regulation for the protection of databases is to be scrutinized as databases make for sizeable amounts of language resources commonly geared toward software-based exploitation.

Corpora used as primary resources may, in numerous occasions, be categorised as ‘original works of authorship’ and consequently be found subject to authorial rights. The latter are prone to a complex intertwining in the context of resources deriving from multiple integrations and contributors. Property rights that are imparted to the authors of a work are the basis for its exploitation, which consists in an exclusive right of representation and distribution of the work. The use or even re-distribution of this work by any third party requires a specific license in agreement with the authors.

With these legal grounds for protection defined, we shall consider if and to which extent the nature and modalities of a specifically linguistic-oriented exploitation of such resources is to require prior cession or licensing of rights. Under French Law, limitations to the author’s rights are introduced to the benefit of educational and/or scientific exploitation of else-wise protected works: whether such educational / scientific justifications cover linguistic finalities is to require specific developments. Insofar authorial rights need to apply, the notion of 'public' will be found to condition the definition and extent of licensed rights.

### 3.2 Existing Licenses

A number of free licenses have been set up to enable the use, study and sharing of specific works, as well as to allow personal additions onto an original work and the dissemination of the resulting work in the same free-oriented spirit. Therefore, free licenses tend to allow an unlimited (free) dissemination and are sometimes not compatible with restrictive or exclusive distribution patterns. This guide identified a number of existing free licenses, showing at the same time their limitations and specific requirements. It is worth observing that although that type of license is called “free”, this does not mean necessarily that the software or any product distributed through such a license can be obtained “free of charge”. One of the best known and most widely used free license is the GNU GPL (GNU's Not Unix General Public License), which was mainly developed for the distribution of software. In France, the CeCILL license was created in order to enable the exchange of software within the research community. Nevertheless, the CeCILL license is not likely to be used beyond a European context, as it provides improved juridical security through reference to the French common law.

Free licenses usually contain a number of authorizations related to minimal or un-existing constraints. For instance, some of the less restrictive licenses, such as the BSD Licence (Berkeley Software Distribution Licence), closely apply to public domain works. As a matter of fact, as highlighted within Creative Commons licenses, free licenses depend on the person who gives the product. Indeed, the main contribution of Creative Commons licenses is that they allow anyone to instinctively draw up his/her own license with his/her own options, instead of looking for several existing licenses to meet one’s needs (‘share what you want, keep what you want’), through a simple yet explicit representation of available options.

Indeed, free licenses are commonly meant for the distribution of software. Interestingly enough, a license has been developed specifically to cover Language Resources: the Lesser General Public License for Linguistic Resources. Building upon the GPL, the LGPLLR introduces interesting specific solutions, but may be found inadequate in the perspective of restrictive distribution patterns.

When dealing with the distribution of Language Resources, we cannot but mention the work carried out by ELDA who offers specific licenses between providers of LRs and users of LRs, in the restricted field of Human Language Technology. The contract between ELDA and the users grants the latter a non-exclusive and non transferable right to use the LRs. Regarding this usage, some providers agree to make their resources available for research and technology/product development, while others only allow distribution for research purposes. As an answer to these different needs, three types of User Licences were drafted:

- *End-User Agreement:* Within this Agreement, the user is engaged in *bona fide* language engineering research activities. The user is not permitted to distribute and market any derivative product or service based on all or a substantial part of the Language Resources.
- *Evaluation Packages End-User Agreement:* Within this Agreement ELDA grants the user the non-exclusive right to use the Evaluation Packages, exclusively for the purposes of evaluating their Human Language Technologies. The user is not permitted to reproduce the Evaluation Packages for commercial or distribution purposes and to commercialise (or distribute for free) in any form or by any means the Evaluation Packages or any

derivative product or services based on all or a substantial part of it.

- *Value-Added Reseller Agreement (VAR):* ELDA grants the user the non-exclusive right to distribute and market any derivative product or service based on all or a substantial part of the Language Resources (according to VAR's commercialization policies).

Further to the above-mentioned licenses, the work of the LDC should also be referred to. Likewise ELDA, LDC considers two types of LR usage, namely research and commercial, under the umbrella of four different membership categories and a non-membership category. The agreements offered for LDC members cover the following types: a) not-for-profit organizations, b) for-profit organizations, c) U.S. government entities, and b) LDC online membership (not-for-profit and government entities). With regard to LDC non-members, a possibility is also offered to acquire data without becoming a member (LDC user agreement for non-members).

# 4. Strategic issues

The aim is to bring into focus how important the sharing of expertise is between the players of the field, in order to help the whole community move forward to technological innovation. Besides, this should prevent us from reiterating the same efforts on similar resources and help optimise the productivity of all organisations involved in this field. More specifically, ELDA has focused its investigation on the BLARK concept (*Basic Language Resource Kit*), which was born from a joint initiative between ELSNET (*European Network of Excellence in Language and Speech*) and ELRA/ELDA. BLARK's objective is to define a minimum set of Language Resources necessary for the development of language technologies and to fill the gaps identified in the field.

Here, ELDA has the chance to take advantage from its multiple experiences in terms of resource promotion, and to a larger extent, in the field of language technology, thanks to its works carried out in partnership with both academic and commercial organisations. In particular, ELDA can make use of its regular information dissemination channels, such as its newsletter, web sites, or its experience at organising international events (LREC conference [18], LangTech [19], evaluation workshops organised in the framework of different conferences or seminars) to support and help advance with these strategic issues.

## 4.1 Sharing Language Resources

Antonio Zampolli, one of the father founders of ELRA, was one of the first to emphasize on the need to associate three language engineering areas: Language Resources, language technologies and applicative projects (Maegaard et al., 2005). In particular, he introduced the need to offer a sound infrastructure to allow a better synergy and coordination of the works within the HLT field. Besides, funding agencies themselves observed this need and now more and more insist on obtaining high justifications on what the European Commission calls "Exit strategy". This "exit strategy" was also enhanced within national programmes. In this regard, the French Technolangue[20] programme is also worth mentioning, whose requirements for the participants were based on:

1) Proposing projects that should bridge the gaps within the HLT field.
2) Making possible the reuse of the results beyond the project itself.
3) Achieving work in synergy between the different submitted projects in order to bring the best expertise possible, sharing knowledge and LRs.

A good example of such implementation can be given through the EVALDA Evaluation Campaigns where several synergies could be set up, through LRs and tools that could be produced for several campaigns inside the project or even disseminated and used for outside follow-up projects.

## 4.2 Filling the gaps: the BLARK

Several activities have been carried out to define a minimal set of LRs to be made available for as many languages as possible, and to map the actual gaps which should be filled in order to meet the needs of the HLT field. Such activities can be gathered under a same concept name, the BLARK (*Basic LAnguage Resource Kit*). To define a minimal set of LRs, two kinds of actions must be taken upstream: the identification of needs with respect to potential Human Language Technologies and the identification of existing LRs. Once the needs and existing LRs have been identified, the following step is to derive a sub-set of items (e.g. tools, data, etc.) that could be considered as priority items for further development. Some priority lists of items have already been identified for a few languages and submitted to large organisations to be developed under external funding. A BLARK initiative was initially designed for the Dutch language by the Dutch Language Union (Cucchiarini et al. 2001a ; 2001b). More recently, new initiatives were set up for the Arabic language through the NEMLAR project (*Network for Euro-Mediterranean LAnguage Resources*), followed by the currently ongoing MEDAR project (*Mediterranean Arabic Language and Speech Technology*)[21]. ELDA also developed an interactive service of the BLARK[22], enabling to identify the needs in terms of LRs with respect to specific applications and corresponding languages.

## 4.3 Capitalizing on Language Resources

Entering the Language Resource market requires a good expertise of this market. Specialised organisations were born in Europe and outside to meet the needs in

[18] http://www.lrec-conf.org
[19] http://www.langtech.it/en/default.htm
[20] http://www.technolangue.net
[21] http://www.nemlar.org
[22] http://www.elda.org/blark

identification, production and LR sharing. Two centres cannot be ignored in this field and are now well known at an international level: ELRA (*European Language Resources Association*) in Europe and LDC (*Linguistic Data Consortium*) in the USA. Both organisations worked at giving access to Language Resources that are produced within national or international projects, as well as LRs produced through individual actions, in order to avoid the loss of such LRs once they are produced, as well as to capitalize on the same LRs beyond their original objectives.

In Asia, GSK [23] (*Gengo Shigen Kyouyuukikou* – Consortium for Language Resources), in Japan, and SITEC [24] (*Speech Information Technology & Industry Promotion Center*) in Korea, are now working at answering the needs of the Asian territory.

## 5. Conclusions

The guide gathers all aspects mentioned previously (technical, legal and strategic), by proposing a course of action to follow in order to produce and make Language Resources available. It covers all steps of production: specifications, standards to follow according to the types of resources and envisaged applications, encoding, storing or exchanging formats, different validation steps. Besides, it defines the means to be used for the distribution of Language Resources, their archiving and maintenance, as well as technical and legal conditions needed for their transfer to third parties. We also mention some constraints and possibilities to enrich a resource or integrate it into another larger resource. Last but not least, it focuses on strategic problems such as the reluctance to share LRs and the promotion of collaboration and exchange between the players of the LR production area.

The guide will be made available through the www.technolangue.net web site. This internet portal, maintained by ELDA, focuses on both written and spoken language technologies. It aims to enhance the HLT field, as well as provide information on its players, highlight and explain the main issues, or inform on the main scientific, technological, industrial and normative evolutions.

## 6. References

Burger S., McLaren V., Yu H. *The ISL Meeting Corpus: The impact on meeting type on speech style*, in Proceedings of ICSLP, Denver, USA, 2002.

Cucchiarini C., Daelemans W. et Strik H., *Strengthening the Dutch Human Language Technology Infrastructure*, ELRA Newsletter Vol. 6 N. 4. 2001a.

Cucchiarini C., Daelemans W. et Strik H., *Strengthening the Dutch Language and Speech Technology Infrastructure*, Actes de la conférence COCOSDA 2001b.

Fersøe H., *Validation Manual for Lexica*, release 2.0, January 2004.

Gandcher F., Hamon H., Mapelli V., Moreau N., Paulsson N., *Réalisation d'un guide de production de ressources linguistiques pour la veille*, ELDA preliminary report, 2008.

Janin A., Baron D. *et al. The ICSI Meeting Corpus*, in Proceedings of ICASSP'03, Hong Kong, China, April 2003.

Maegaard B., Choukri K., Calzolari N., Odijk J., *ELRA – European Language Resources Association - Background, Recent Developments and Future Perspectives*, LRE Journal, Volume 39, Number 1 / February, 2005.

McEnery T., Burnard L., Wilson A. and Baker, *Validation of Linguistic Corpora*, 28 April 1998.

Moreau N. *et al. Exploitation Material for the CHIL Evaluation Campaign 3*. CHIL Public Deliverable D7.14, 2007a.

Moreau N., Mostefa D., Stiefelhagen R. *Perceptual Component Evaluation and Data Collection*, in: Alex Waibel, Rainer Stiefelhagen (Eds.), "CHIL: Computers in the Human Interaction Loop", Springer, 2007b.

Van den Heuvel H., Boves L., Sanders E., *Validation of Content and Quality of Existing SLR : Overview and Methodology*, Deliverable 1.1, 21 January 2000.

[23] GSK: http://www.gsk.or.jp/index_e.html

[24] SITEC: http://www.sitec.or.kr