Review on the Existing Language Resources for Languages of France

Thibault Grouas[1], Valérie Mapelli[2], Quentin Samier[2]
[1] Délégation générale à la langue française et aux langues de France
6 rue des Pyramides, 75001 Paris, France
Email: thibault.grouas@culture.gouv.fr
[2] ELDA/ELRA
9 rue des Cordelières, 75013 Paris, France
Email: mapelli@elda.org; quentin@elda.org

Abstract

With the support of the DGLFLF, ELDA conducted an inventory of existing language resources for the regional languages of France. The main aim of this inventory was to assess the exploitability of the identified resources within technologies. A total of 2,299 Language Resources were identified. As a second step, a deeper analysis of a set of three language groups (Breton, Occitan, overseas languages) was carried out along with a focus of their exploitability within three technologies: automatic translation, voice recognition/synthesis and spell checkers. The survey was followed by the organisation of the TLRF2015 Conference which aimed to present the state of the art in the field of the Technologies for Regional Languages of France. The next step will be to activate the network of specialists built up during the TLRF conference and to begin the organisation of a second TLRF conference. Meanwhile, the French Ministry of Culture continues its actions related to linguistic diversity and technology, in particular through a project with Wikimedia France related to contributions to Wikipedia in regional languages, the upcoming new version of the "Corpus de la Parole" and the reinforcement of the DGLFLF's Observatory of Linguistic Practices.

Keywords: regional languages of France, inventory, language resources.

1. Context and Aim of the Project

As a result of a partnership with the *Délégation générale à la langue française et aux langues de France* (DGLFLF, French Ministry of Culture and Communication) [1], the Evaluations and Language resources Distribution Agency (ELDA) [2] conducted an inventory of language resources currently existing for the regional languages of France throughout mainland France and French overseas departments and territories (Leixa et al., 2014). The primary goal of the inventory was to assess the exploitability of these identified resources within different kinds of language technologies. With adapted technologies, regional languages are expected to gain more visibility and applicability among a wider audience.

For that purpose, the task was to identify the main channels of production and dissemination and to provide a non-exhaustive list of the existing language resources.

The following section will explain in more detail how the task was carried out and how the scope of our study was defined by laying down a number of criteria. The next section will fully describe the outcomes of the whole project. The third and final section will provide the implications, recommendations and future prospects of the study.

2. Up-front Work and Methodology

The scope of the identification task had to be defined in order to ensure the accomplishment of the assigned objectives mentioned in the section above.

First of all, we agreed on the types of language technologies most adapted to our goal, based upon the list provided by the MetaNet White Paper [3] (Mariani et al., 2010). In particular, three of them caught our attention, considered as representative for the processing and further dissemination of regional languages: automatic translation, voice recognition/synthesis and spell checkers.

The scope was also defined based on the classification of languages. Two sources of information helped determine it: the DGLFLF institution (French Ministry of Culture and Communication) (Délégation générale à la langue française et aux langues de France, 2010) and the Ethnologue website. With the help of these two sources, we based our study on a range of 84 languages spoken in France and its overseas departments and territories. Those include the languages identified by the French Ministry of Culture and Communication, however with a different classification which may be debatable and reviewed according to linguistic community standards: for instance, Occitan languages are regarded as one single language by the DGLFLF whereas they are seen as individual languages by Ethnologue specialists. It is important, therefore, to consider each of these sources of information as biased and subjective viewpoints on these sociolinguistic realities, bound to be multiple. At the time of the review, it appeared to be the most convenient solution for the identification task.

In order to classify the identified languages, three criteria were applied using the information from the Ethnologue website. These criteria include:

- the different language families;
- the number of speakers;
- the transmission modalities (oral, written or signed).

Specific criteria were adopted to determine the type of information to describe language resources, but also different sources of information that are likely to provide language resources, such as newspapers, radio channels, institutional or cultural sites. Processing them electronically will enable us to produce different types of language resources (corpora, lexica, spoken or even multimodal resources) (Gandcher et al., 1998). Such

[1] http://www.culturecommunication.gouv.fr/Politiques-ministerielles/Langue-francaise-et-langues-de-France

[2] http://www.elda.org/

[3] http://www.meta-net.eu/whitepapers

criteria are inspired from the metadata in use in the ELRA[4] and Linguistic Data Consortium (LDC)[5] catalogues, as well as the OLAC nomenclature[6]. For the language resources, the criteria consist of the type of the language resource identified (written corpus, spoken corpus, parallel corpus, multimedia resource, lexicon, grammar, thesaurus), its name, the related language(s), the description of the resource, its volume, existing or potential applications that can be developed with the resource, its location on the Internet, the provider(s), its availability and the possible rights associated to its usage. As far as the sources are concerned, an ontology was defined to retrieve the following information: name, description, URL, related language(s), possible applications and contact details needed for locating those resources.

As a result of the identified classification and defined metadata, an investigation was carried out and a resulting inventory was made of the existing language resources and of different sources of information. For the identification of language resources, we retrieved most of them from the main channels of dissemination: the ELRA catalogue, LRE Map[7], Meta-Share initiative[8], the LDC catalogue, the OLAC initiative. Beyond those main channels, other channels of information were exploited. In particular, we would like to mention the participation of Lo Congrès[9], the interregional organisation for the regulation of the Occitan language: they provided a significant list of resources for the Occitan language that were not identified from the main channels. Moreover, we can also mention the work carried out within the "Corpus de la Parole" programme[10], funded by the French Ministry of Culture and Communication, which provided over 2,000 hours of audio data.

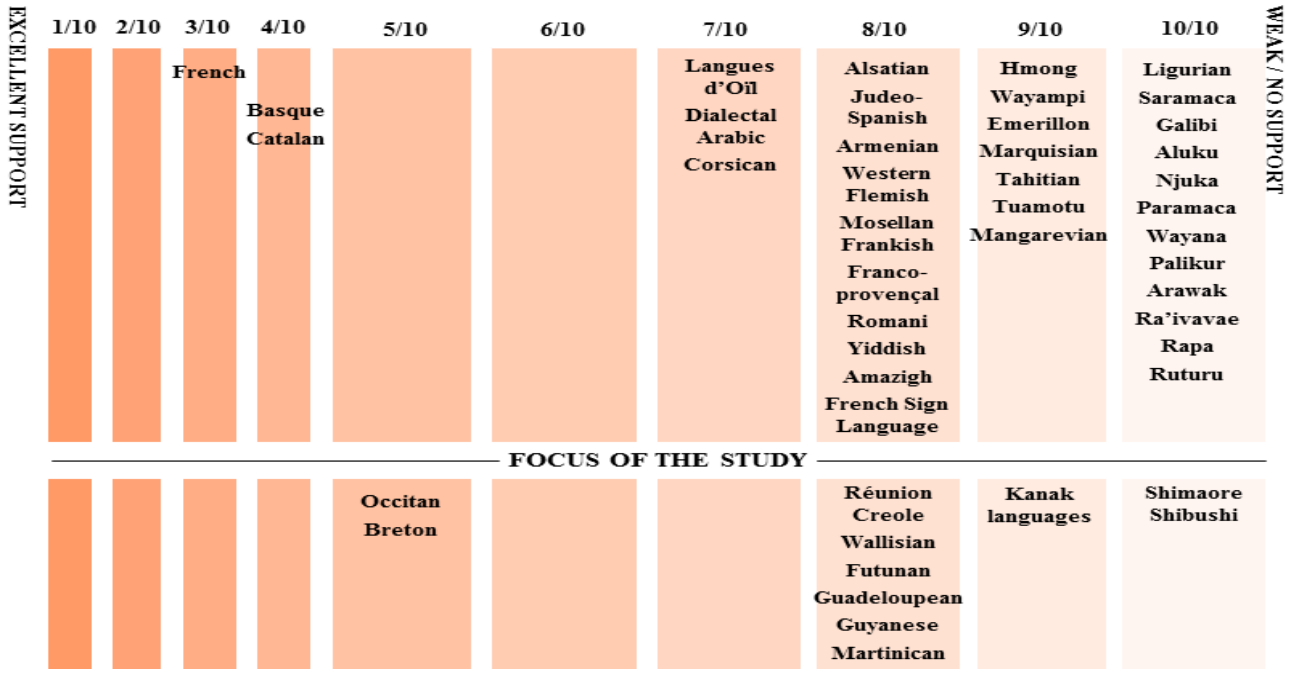
Given the substantial volume of the information retrieved from the Internet and the catalogues, the question arose as to how and in what format the inventory should be presented; this was resolved by our decision to create a MySQL database where information may be compiled with corresponding statistics, and enriched at a later stage. The database is divided into two groups, the sources group and the language resources group built on the basis of the metadata defined above.

3. Results

A total of 2,299 Language Resources were identified within the study. They are subdivided as follows: 1,417 spoken corpora, 425 written corpora, including parallel corpora, 181 lexica, 206 multimedia/multimodal corpora, 16 grammars/language models, 1 ontology, 7 thesauri/wordnets, 17 media (newspapers) collections, 19 TV/Radio resources, 10 mixed corpora, i.e. combining several types of LRs.

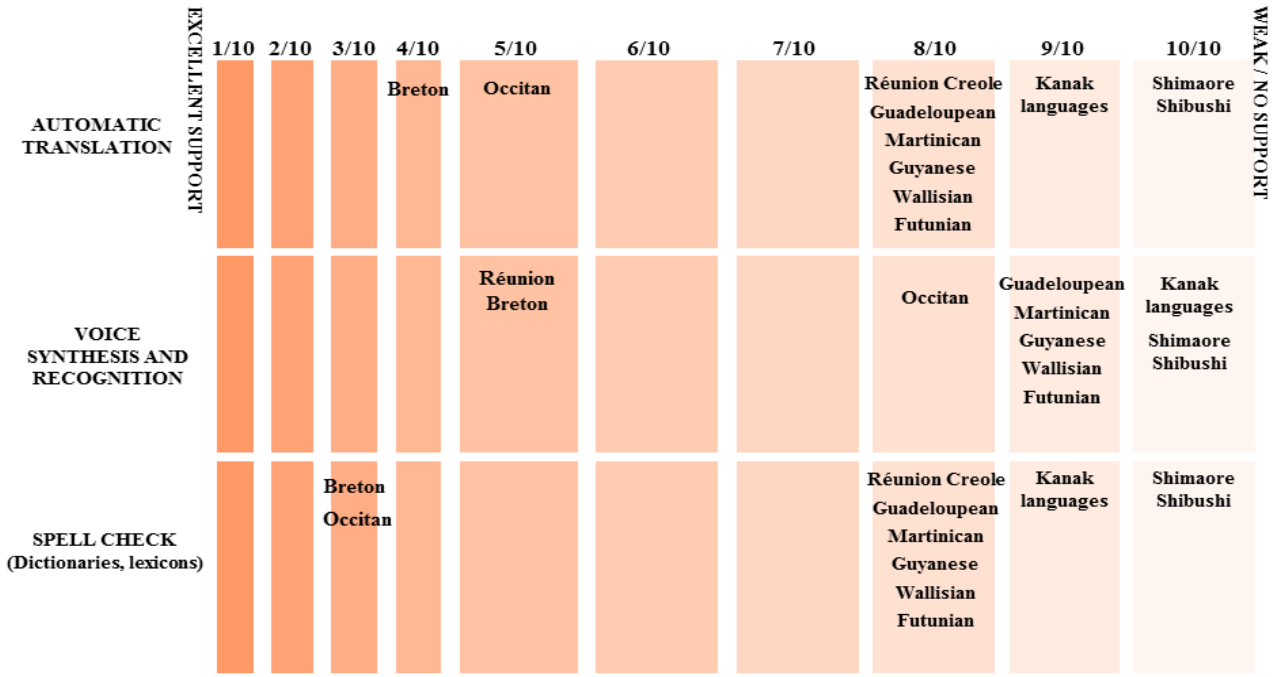
Among the ten most represented languages in the report are, in first position, Occitan (with 669 identified resources), followed by Breton (with 450 resources). The Catalan language, even with 47 resources, still stands among the ten most represented languages.

Due to the high number of languages to be dealt with in a short period of time and in consultation with the DGLFLF, we decided to focus on three language groups: overseas languages, Breton and Occitan. Such a decision also impacted the inventory itself, in which we could regretfully notice the lack of results for some languages, like Corsican.



Graph 1: Representation of languages in terms of language resources

[4] http://catalog.elra.info
[5] https://catalog.ldc.upenn.edu
[6] http://www.language-archives.org
[7] http://www.resourcebook.eu
[8] http://www.meta-share.eu
[9] http://www.locongres.org
[10] http://cocoon.huma-num.fr



Graph 2: Representation of languages in terms of technologies

This could however be done at any subsequent stage of the inventory. Furthermore, our wish was not to focus only on resources, but also on how they may actually be used with language technologies. With regard to the three language groups taken into account, we focussed on three technologies considered as most relevant for the processing and further dissemination of regional languages, i.e. automatic translation, voice recognition/synthesis and spell checkers, in order to analyse the feasibility of developing such technologies with regard to those three language groups.

The results of the analysis were gathered in the two graphs presented herein, inspired from the MetaNet White Paper. These show the position of the languages with respect to their high or low representation in terms of language resources and applications (on a scale from 1 to 10 where 1 is the highest and 10 the lowest).

A report has been published and is now available on both the ELRA/ELDA[11] and DGLFLF websites[12].

The inventory is freely downloadable as a spreadsheet (.xlsx and .ods)[13]. Access to the MySQL database will be provided at a later stage. It includes on the one hand a "Sources" section that lists the main available media (mostly news media) and on the other hand a "Resources" section that lists the language corpora that can be used within Natural Language Processing (NLP) systems.

4. Following-up from the Review

4.1. TLRF2015 Conference

To move forward with the first outcomes of this study, among others [14], ELDA, the IMMI-CNRS, DGLFLF, LIMSI-CNRS and ORTOLANG cooperatively organised a two-day conference presenting the state of the art in the field of the Technologies for Regional Languages of France (TLRF)[15] on 19 and 20 February 2015, in the Paris area. The participants – numbering around eighty – were mostly linguists, NLP specialists, representatives of the State's national and regional authorities and public offices of regional languages. The aim was to:

- conduct a survey on the development of existing technologies;
- show successful examples of development for some languages;
- propose realistic solutions that can overcome the existing gaps.

This conference drew its origin from the observation that breakthroughs in NLP had been achieved, but only for 1%

[11] For general information, see: http://www.elra.info/en/projects/archived-projects/review-existing-lrs-france/ and for the downloadable report, see: http://www.elra.info/media/filer_public/2014/12/17/rapport_dglflf_05112014-1.pdf

[12] http://www.culturecommunication.gouv.fr/Politiques-ministerielles/Langue-francaise-et-langues-de-France/Politiques-de-la-langue/Langues-et-numerique/Les-technologies-de-la-langue-et-la-normalisation/Inventaire-des-ressources-linguistiques-des-langues-de-France

[13] http://www.elra.info/media/filer_public/2015/02/18/inventaire_05112014.xlsx or http://www.elra.info/media/filer_public/2015/02/18/inventaire_05112014.ods

[14] See for example, *Étude sur la place des langues de France sur Internet*, Réseau Maaya and Délégation générale à la langue française et aux langues de France, http://www.culturecommunication.gouv.fr/Politiques-ministerielles/Langue-francaise-et-langues-de-France/Politiques-de-la-langue/Langues-et-numerique/La-diversite-linguistique-et-la-creation-artistique-dans-le-domaine-numerique/Etude-sur-la-place-des-langues-de-France-sur-l-internet

[15] http://tlrf2015.sciencesconf.org/

of the languages spoken in the world, with little to no coverage for most regional languages, even in France. Some languages, for example Basque and Catalan in Spain, thrive better due to the combination of a strong political will with scientific and technical knowledge. These languages therefore benefit from more rapidly developed state-of-the-art technologies and are thus equipped with a larger number of language resources and tools. This favourable situation allows for an increase in basic research activities on those languages and for a faster and more efficient rise to the grand challenge of implementing a real multilingualism acknowledging a variety of regional languages.

The programme of this first conference consisted of oral presentations and panel sessions that should later result in drafting an action plan, backed up by a series of future conferences, if needed. After a broad overview of the latest developments of the languages of France and a general presentation of the challenges posed by language technologies, the goals were set out of several research programmes for collecting resources and tools, such as the "Corpus de la parole" programme or the "Restaure" project. The representativeness issue of the languages of France in the digital field was tackled, in particular with respect to the studies carried out on their electronic processing, but also their presence on the Internet, especially on Wikipedia. Some participants discussed the possibility of sharing resources for linguistically related French creoles while others talked about speech processing and translation for underequipped languages.

On the second day of the conference, round tables were organised that brought debates on major challenges and issues of the vast field of language technology to a broader public. However, despite the first successful outcomes, we are aware that there is still a substantial amount of work to be done to advance this field, fuelling a growing need for further developments in fundamental and applied research. The whole conference was broadcast live on the internet and the recorded videos are now viewable online[16].

4.2. Current Actions from the DGLFLF

The proceedings of the TLRF conference are being edited and will be published by the end of the year. We have already gathered all papers from contributors and speakers. These proceedings will prove a valuable reference because they represent a significant effort towards the development of technologies for the promotion of linguistic diversity.

The next step will be to activate the network of specialists built up during the TLRF conference and to begin the organisation of a second TLRF conference, with the help of regional agencies and associations. This conference, which should be organised by regional organisations, could focus on establishing or choosing a platform for publishing language resources and data as well as finding better ways of structuring new research or industrial projects for the technological development of the languages of France.

Meanwhile, the French Ministry of Culture continues its actions related to linguistic diversity and technology. Among these actions, an ongoing project with *Wikimédia France* related to contributions to Wikipedia in regional languages will lead to a conference on this issue in January 2016 and to the release of new applications facilitating the uploading of recorded speech in regional languages to Wikimedia platforms (Wikipedia, Wiktionary, Wikimedia Commons, etc.).

A new version of the "Corpus de la Parole" website will be released in 2016, focussing on the re-usability of data with linked open data standards, such as RDF, and the availability of a free-to-use triple store.

Finally, the Observatory of Linguistic Practices (*Observatoire des pratiques linguistiques*) of the DGLFLF will be reinforced and will provide a new online platform to gather all information regarding languages of France (studies, data sets, standards, resources, regional languages related websites, geo-linguistic maps and representations, etc.).

5. Bibliographical References

Europe's Languages in the Digital Age (2012). META-NET White Paper Series, Springer.

Délégation générale à la langue française et aux langues de France (2010). Les langues de France, Références (http://www.culturecommunication.gouv.fr/content/download/93669/841697/version/3/file/ref_2010_lg_de_France_def.pdf)

Leixa J., Mapelli V., Choukri K. (2014). Inventaire des ressources linguistiques des langues de France.

Gandcher, F., Hamon, O., Mapelli, V., Moreau, N., Paulsson, N., Mostefa, D. (1998). Réalisation d'un guide de production de ressources linguistiques pour la veille.

Mariani, J., Paroubek, P., Francopoulo G., Max A., Yvon F., Zweigenbaum P. (2012). La langue française à l'ère du numérique.

Ramchetty, R. (1998). Rapport sur l'état de la coopération régionale.

[16] https://webcast.in2p3.fr/events-tlrf