

Language Technologies for Regional Languages of France: The RESTAURE Project

Delphine Bernhard¹, Myriam Bras², Pascale Erhart¹,

Anne-Laure Ligozat³, Marianne Vergez-Couret⁴

¹LiLPa, Université de Strasbourg, France, ²CLLE, Université de Toulouse, CNRS, UT2J, France,

³LIMSI, CNRS, ENSIIE, Université Paris-Saclay, F-91405 Orsay, France, ⁴ FoReLLIS, Université de Poitiers, France

¹{dbernhard,pascale.erhart}@unistra.fr, ²myriam.bras@univ-tlse2.fr,

³anne-laure.ligozat@limsi.fr, ⁴marianne.vergez.couret@univ-poitiers.fr

Abstract

The RESTAURE project (2015-2018) aimed at providing digital resources and natural language processing (NLP) tools for three regional languages of France: Alsatian, Occitan and Picard. These languages belong to different language families and are characterized by heterogeneous sociolinguistic situations. In this paper, we focus on the main challenges faced during the project and detail the solutions that we have implemented for the development and distribution of the resources and tools produced. We also present the main lessons learned from the RESTAURE project.

Keywords: Alsatian, Occitan, Picard, language technologies

Résumé

Le projet RESTAURE (2015-2018) visait à fournir des ressources numériques et des outils de traitement automatique des langues (TAL) pour trois langues régionales de France : alsacien, occitan et picard. Ces langues appartiennent à des familles linguistiques différentes et se caractérisent par des situations sociolinguistiques hétérogènes. Dans cet article, nous nous concentrons sur les principaux défis rencontrés au cours du projet et détaillons les solutions que nous avons mises en œuvre pour le développement et la distribution des ressources et outils produits. Nous présentons également les principaux enseignements tirés du projet RESTAURE.

1. Introduction

France has only one official language, French, but many more regional languages are present on the French metropolitan territory (23 according to Leixa et al. (2014), but there is no consensus on this number). In contrast to French, these regional languages are poorly equipped with linguistic resources and NLP tools. In this article, we present the results of the RESTAURE project¹ (2015-2018) aimed at providing digital resources and natural language processing (NLP) tools for three regional languages of France: Alsatian, Occitan and Picard. It brought together researchers from four French research units located in Strasbourg (Université de Strasbourg – LiLPa), Toulouse (Université Toulouse Jean-Jaurès – CLLERSS), Amiens (Université de Picardie Jules Verne – Habiter le monde) and Orsay (LIMSI).

We will first briefly describe the three regional languages of France included in the project (Section 2.). We will then present some challenges to providing language technologies for these languages (Section 3.). We will also discuss the solutions, based on recent recommendations to improve digital language vitality of under-resourced and minority languages (Soria et al., 2013; Ceberio Berger et al., 2018) (Section 4.). Finally, we will present the main lessons learned from the RESTAURE project (Section 5.).

2. Description of Alsatian, Occitan and Picard

2.1. Alsatian

The Germanic Alsatian dialects are spoken in the North-East of France. The dialectal domain of High-German dialects in France actually stretches farther than the former Alsace region and encompasses part of the Moselle department. Moreover, the dialectal domain can be decomposed in several areas, with Low Alemmanic, High Alemmanic and Central German Franconian dialects being represented. The Alsatian dialects can be traced back to the 6th century and the linguistic changes brought by the Alemanni and the Franks (Huck, 2015). The last decades have however seen a decline in the use of the Alsatian dialects, with French being used as the main language of communication in the region.

The Alsatian dialects have mainly been used orally, with a small literary production since 1816 (mainly theater plays and poetry). Spelling is not standardized, which, in addition to *spatial variation* (both on the phonological and lexical levels) accounts for the very diverse graphical variants found in writing.

2.2. Occitan

Occitan is a romance language spoken in southern France and in Val d’Aran in Spain and in several valleys of Italy. Occitan has several varieties organized in dialects. The most accepted classification suggested by Bec (1996) includes Auvergnat, Gascon, Languedocien, Limousin, Provençal and Vivaro-Alpin. However, each dialect has

¹<http://restaure.unistra.fr/>

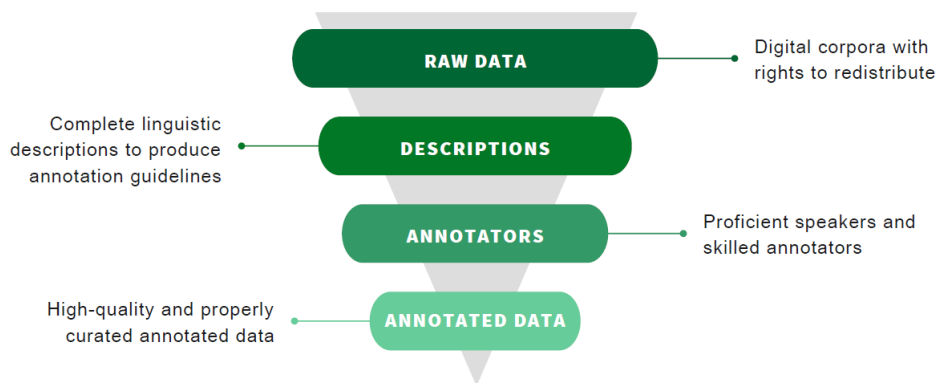


Figure 1: Data bottleneck.

also internal variations. Occitan is written since the Middle Ages and an extensive body of literature has been produced. Although much less socialised than it was before World War II, Occitan is now present in newspapers, on the Internet, on the radio and television, and in some schools and universities.

There are two main spelling standards: the ‘mistralienne’ spelling designed in the mid-19th by Frederic Mistral and the ‘classical’ spelling from the 20th century based on medieval conventions whose aim is to minimize the dialectal differences while keeping dialectal particularities (Sibille, 2006). However, literature in Occitan is characterized by a plethora of non-standard individual spellings.

2.3. Picard

Picard is a langue d’oil (Romance language group) spoken in the North of France (Hauts-de-France) and the Belgian province of Hainaut. Picard has several varieties and spelling is not standardized. Picard is, however, used in writing, as shown by the PICARTEXT database (Eloy et al., 2015), which includes literary works, totalling about 5 million words.

3. Challenges

In this section, we present the most important challenges we have faced during the project.

3.1. Data Bottleneck

Figure 1 sums up what we call the *data bottleneck* challenge for collecting and producing high-quality and properly curated linguistically annotated data. Even if the problems presented are not confined to under-resourced languages, they are even more important for them.

First, collecting raw corpora is made difficult by the scarcity of available resources. For instance, it is usually easy to collect very large corpora on the Web (e.g. using Wikipedia) for languages with many speakers and a good online presence. This is much more of a challenge for under-resourced languages.

Second, accurate and complete linguistic descriptions are needed to enrich corpora with annotations (e.g., part-of-speech, morphosyntactic features). Up-to-date grammars are very difficult to find for regional languages of France: if they exist, they are often outdated or incomplete.

Third, concerning the annotation work *per se*, it is hard to recruit people who are both proficient speakers and skilled annotators.

3.2. Dialectal and Spelling Variation

As already hinted at in Section 2., Alsatian, Occitan and Picard are neither homogeneous nor fully standardized. Different varieties or dialects of these regional languages can be identified in each region. Spelling conventions are either rather recent, or not much used, or even accommodate for dialectal particularities. All in all, dialectal and spelling variation is challenging for NLP tools. For instance, the uncontrolled use of punctuation marks makes it difficult to develop reliable tokenizers, which automatically break down texts into words (Bernhard et al., 2017)

4. Solutions

The solutions we have implemented include a large part of the recommendations by Soria et al. (2013), both for the development of resources and tools (see Section 4.1.) and their distribution (see Section 4.2.).

4.1. Development of Resources and Tools

The development of resources and tools was based on three main principles, in accordance with (Soria et al., 2013): (1) cooperation, (2) use of standards and (3) re-use and recycling of existing tools.

4.1.1. Cooperation

Cooperation between the teams involved in the RESTAURE project was an important asset, all the more so as the different teams specialized in different domains and had variable previous experience in producing language technologies for under-resourced languages. It led to the development of a common corpus annotation workflow (see Figure 2) and to collaboration in carrying out the various sub-tasks.

For instance, Strasbourg and Toulouse cooperated to perform Optical Character Recognition (OCR) for corpus acquisition (Vergez-Couret et al., 2015). Strasbourg and Amiens worked together to develop a tokenizer for Picard (Bernhard et al., 2017). Orsay provided help to Amiens to parse and format lexicons. Finally, Orsay and Strasbourg

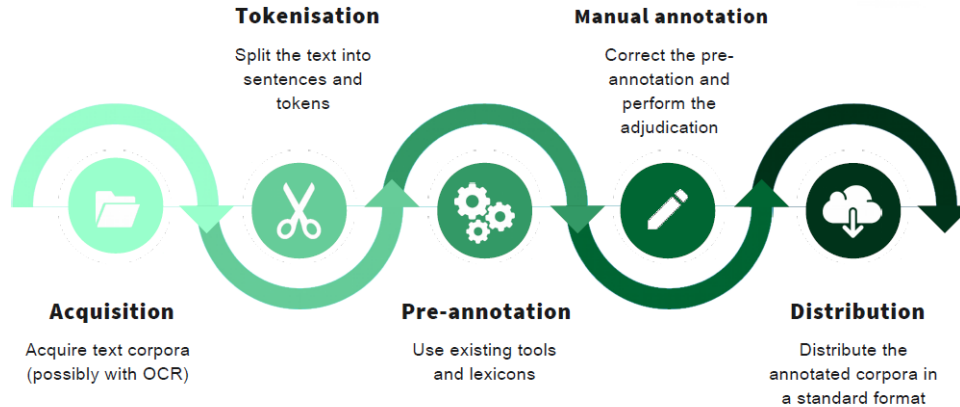


Figure 2: Corpus annotation workflow. Icons made by Tomas Knop, Smashicons, Freepik from www.flaticon.com

cooperated on the task of identifying place names for Alsatian (Bernhard et al., 2018b). Clearly and in retrospect, many tasks could not have been accomplished, or in a less sophisticated form, without the collaboration between the different teams. This cooperation made it possible to compensate, to some extent, for the lack of human resources and specialists for the regional languages under study.

4.1.2. Use of Standards

As Soria et al. (2013) write:

“Use of standards is the key to interoperability of resources, as they allow resource sharing, re-usability, maintainability and long-term preservation.”

We thus chose to share the annotated corpora produced by the RESTAURE project in the CONLL-U format, defined in the *Universal Dependencies* (UD) project (Nivre et al., 2016). This format is directly usable for training POS (Part-Of-Speech) tagging tools such as spaCy² or UDPipe (Straka and Straková, 2017). Moreover, the Universal POS tags defined in UD helped us define tagsets for Alsatian and Picard as well as write annotation guidelines based on the UD recommendations. The original tagsets for Alsatian, Occitan and Picard are not strictly identical to the UD POS tags, but could be transformed into these tags using a correspondence table. The procedure for transforming our corpora into UD format is described in (Miletic et al., 2019).

4.1.3. Re-Use and Recycling of Existing Tools

During the course of the project, we re-used and recycled existing tools, whenever possible:

- OCR (Vergez-Couret et al., 2015):
 - Tesseract (Smith, 2007)
 - Jochre (Urieli and Vergez-Couret, 2013)
- Part-of-speech (POS) tagging (Vergez-Couret and Urieli, 2015; Bernhard et al., 2018a):
 - for Occitan, Talismane (Urieli, 2013) and APERTIUM (Armentano I Oller, 2008)
 - for Alsatian, TreeTager for German (Schmid, 1994)

- Corpus annotation (Bernhard et al., 2018a): Analog tool (Lay and Pincemin, 2010)

4.2. Distribution of Resources and Tools

The distribution of resources and tools produced during the RESTAURE project also followed three main principles, again in accordance with (Soria et al., 2013): (1) document resources and technologies, (2) be open and (3) share and sustain. The outputs of the RESTAURE project are shared on the Zenodo platform (<https://zenodo.org/communities/restaure>), under a Creative Commons Attribution Share Alike 4.0 licence (CC-BY-SA). The resources and tools are associated with a DOI and are fully documented.

5. Lessons Learned from the RESTAURE Project

Finally, we detail the lessons learned during the course of the project:

Cooperation is key This work on regional languages of France could not have been carried out without real cooperation between various teams with complementary skills (sociolinguistics, dialectology, natural language processing). The parallel work on several languages made it possible to benefit from the experiences carried out on other languages and thus gain in efficiency. The problems that arose in one language led to increased vigilance on this subject in the other languages.

Do not feel inferior to “big” languages Working on under-resourced languages often means starting building language technologies from (or almost from) scratch. It is easy to feel that you are far behind in comparison to better-resourced languages with many more researchers, resources and tools. Producing language resources requires time and the means to do so, and both are rare for under-resourced languages. These extrinsic constraints are difficult to control but should not undermine the desire of researchers to keep working on these languages. This requires that funding agencies as well as program and reviewing committees acknowledge the specific challenges of work on under-resourced languages.

²<https://spacy.io/>

Do not reinvent the wheel This is an important principle. First, this means that instead of developing new tools, it is often less time-consuming to try and find a similar tool which can be adapted to your own needs. It is also necessary to learn from similar projects, including e.g. existing guidelines for annotating corpora. Within the project, participants should share a common workflow and use the same tools, if possible. In return, it is important to distribute the resources that have been created, so that the work is beneficial to others.

Focus on data rather than tools Nowadays, most NLP tools are able to learn from data. Methods have evolved from being predominantly based on rules towards machine learning techniques, which are in principle applicable to a wide variety of languages. The main condition is that data are available for re-training them. It is therefore advisable to concentrate on data collection and annotation, rather than on the development of tools. As stressed earlier, tools can then be re-used or re-trained.

6. Acknowledgements

This work was supported by the French “Agence Nationale de la Recherche” (ANR) through the RESTAURE project (no.: ANR-14-CE24-0003).

7. Bibliographical References

- Armentano I Oller, C. (2008). Traduction automatique occitan-catalan et occitan-espagnol: difficultés affrontées et résultats atteints. In *IXème Congrès International de l'Association Internationale d'Etudes Occitanes*, Aachen.
- Bec, P. (1996). *La langue occitane*. Paris, PUF.
- Bernhard, D., Todirascu, A., Martin, F., Erhart, P., Steiblé, L., Huck, D., and Rey, C. (2017). Problèmes de tokénisation pour deux langues régionales de France, l’alsacien et le picard. In *Actes de l’atelier “Diversité Linguistique et TAL” – DiLiTAL 2017*, pages 14–23.
- Bernhard, D., Ligozat, A.-L., Martin, F., Bras, M., Magistry, P., Vergez-Couret, M., Steiblé, L., Erhart, P., Hathout, N., Huck, D., Rey, C., Reynés, P., Rosset, S., Sibille, J., and Lavergne, T. (2018a). Corpora with Part-of-Speech Annotations for Three Regional Languages of France: Alsatian, Occitan and Picard. In *Proceedings of the 11th edition of the Language Resources and Evaluation Conference*, Miyazaki, Japan, May.
- Bernhard, D., Magistry, P., Ligozat, A.-L., and Rosset, S. (2018b). Resources and Methods for the Automatic Recognition of Place Names in Alsatian. In Andrew U. Frank, et al., editors, *Corpus-Based Research in the Humanities*, volume 1 of *Proceedings of the Second Workshop on Corpus-Based Research in the Humanities CRH-2*, pages 35–44, Vienna, Austria.
- Ceberio Berger, K., Gurrutxaga Heraiz, A., Baroni, P., Davyth, H., Kruse, E., Quochi, V., Russo, I., Salonen, T., Sarhimaa, A., and Soria, C. (2018). Digital Language Survival Kit. The DLDP Recommendations to Improve Digital Vitality. Technical report.
- Eloy, J.-M., Martin, F., and Rey, C. (2015). PICARTEXT: Une ressource informatisée pour la langue picarde. In *Actes de TALaRE 2015 - Traitement Automatique des Langues Régionales de France et d'Europe*.
- Huck, D. (2015). *Une histoire des langues de l'Alsace*. La Nuée bleue, Strasbourg. 24 cm. Bibliogr. p. 447-457.
- Lay, M.-H. and Pincemin, B. (2010). Pour une exploration humaniste des textes: AnaLog. In *Proceedings of 10th International Conference Journées d'Analyse statistique des Données Textuelles*.
- Leixa, J., Mapelli, V., and Choukri, K. (2014). Inventaire des ressources linguistiques des langues de France. Technical Report ELDA-DGLFLF-2013A.
- Miletic, A., Bernhard, D., Bras, M., Ligozat, A.-L., and Vergez-Couret, M. (2019). Transformation d’annotations en parties du discours et lemmes vers le format Universal Dependencies : étude de cas pour l’alsacien et l’occitan. In *Actes de TALN 2019*.
- Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., and others. (2016). Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*.
- Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49.
- Sibille, J. (2006). L’occitan, qu’es aquò. *Langues et Cité : bulletin de l’observation des pratiques linguistiques*, 10.
- Smith, R. (2007). An overview of the Tesseract OCR engine. In *Proceedings of the 9th International Conference on Document Analysis and Recognition*, pages 629–633.
- Soria, C., Mariani, J., and Zoli, C. (2013). Dwarfs sitting on the giants’ shoulders—how LTs for regional and minority languages can benefit from piggybacking major languages. In *Proceedings of XVII FEL Conference*.
- Straka, M. and Straková, J. (2017). Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*.
- Urieli, A. and Vergez-Couret, M. (2013). Jochre, océrisation par apprentissage automatique : étude comparée sur le yiddish et l’occitan. In *Actes de TALARE 2013 : Traitement Automatique des Langues Régionales de France et d'Europe*, pages 221–234.
- Urieli, A. (2013). *Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. Ph.D. thesis, Université de Toulouse II le Mirail.
- Vergez-Couret, M. and Urieli, A. (2015). Analyse morphosyntaxique de l’occitan languedocien : l’amitié entre un petit languedocien et un gros catalan. In *Actes de TALARE 2015 : Traitement Automatique des Langues Régionales de France et d'Europe*.
- Vergez-Couret, M., Bernhard, D., Urieli, A., Bras, M., Erhart, P., and Huck, D. (2015). Océrisation de textes pour les langues régionales. Regards croisés sur l’occitan et l’alsacien. In Emmanuelle Chevre Pébayle, editor, *Actes du 10e colloque ISKO France 2015*, pages 250–269.