

Le projet RESTAURE

Delphine Bernhard, Marianne Vergez-Couret

▶ **To cite this version:**

Delphine Bernhard, Marianne Vergez-Couret. Le projet RESTAURE. Colloque sur les technologies pour les langues régionales de France (TLRF 2015), Délégation générale à la langue française et aux langues de France, laboratoire de recherche en informatique pluridisciplinaire (LIMSI) - Centre national de la recherche scientifique (CNRS), Institut des technologies multilingues et multimédias de l'information (IMMI), Feb 2015, Meudon, France. pp.96-100. hal-01297835

HAL Id: hal-01297835
https://hal.science/hal-01297835

Submitted on 5 Apr 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Le projet RESTAURE

Delphine Bernhard, Laboratoire linguistique, langues, parole (LiLPa) – université de Strasbourg
Marianne Vergez-Couret, Laboratoire Cognition, Langues, Langages, Ergonomie (CLLE) - Équipe de recherche en synthaxe et en sémantique (ERSS) – université de Toulouse II Jean Jaurès

Transcription revue et complétée avec la participation de Myriam Bras, CLLE-ERSS – université de Toulouse II Jean Jaurès et de Christophe Rey, Linguistique Et Sociolinguistique : Contacts, Lexique, Appropriations, Politiques (LESCLAP) (CERCLL-EA 4283), université de Picardie Jules Verne

Le projet Ressources Informatisées et traitement automatique pour les langues régionales (RESTAURE) est un projet financé par l'ANR, entamé au mois de janvier 2015 pour une durée de 42 mois. Il comporte trois objectifs principaux :

- acquisition et normalisation de ressources (corpus et lexiques) ;
- développement d'outils pour l'acquisition et l'analyse de corpus ;
- diffusion des résultats auprès du grand public.

Les langues régionales de France concernées par le projet sont au nombre de trois : le picard, l'alsacien et l'occitan. Chacune de ces langues est représentée par un laboratoire partenaire : LESCLAP à Amiens pour le picard, LiLPa à Strasbourg pour l'alsacien, et CLLE-ERSS à Toulouse pour l'occitan. À cela s'ajoute un laboratoire en région parisienne, le LIMSI-CNRS, qui travaille sur les aspects de traitement automatique des langues.

La motivation principale du projet est le manque de ressources informatisées pour les langues régionales de France, en particulier pour les trois langues concernées par le projet.

1. État des lieux des ressources et outils existants

Pour ce qui est des corpus, la langue la plus avancée des trois est le picard, car il existe une base textuelle appelée Picartext, que nous présenterons plus

en détails par la suite. L'occitan est aussi relativement bien avancé, grâce à plusieurs projets en cours, dont la construction de la base textuelle BaTelÒc, que nous présenterons également, alors qu'il n'existe aucun corpus à l'heure actuelle pour l'alsacien : il s'agit là d'une lacune que nous souhaiterions combler.

Pour ce qui est des lexiques pour le Traitement Automatique des Langues et l'étiquetage morpho-syntaxique, quelques travaux ont été réalisés pour l'alsacien et l'occitan (Bernhard, 2014 ; Bernhard et Ligozat, 2013 ; Vergez-Couret et Urieli, 2014), mais il faut reconnaître que l'on n'atteint pas encore des niveaux de performance similaires à ceux du français. Les lacunes sont encore plus importantes pour l'analyse syntaxique et la lemmatisation.

Différentes raisons expliquent ce manque de ressources, mais le défi majeur est celui de la variation graphique, qui peut poser des problèmes aux outils automatiques.

En alsacien, l'exemple du mot « lundi » est un cas intéressant : on peut le trouver sous les formes « Mantig », « Mandig », « Mandi », « Mändàach », « Mändàà », « Mondàà ». Ces diverses formes ont peu de caractères en commun, mais constituent tout de même des variantes d'une même unité lexicale et devraient donc être reconnues comme telles. Il s'agit là du défi scientifique majeur auquel le projet RESTAURE va tenter d'apporter des solutions.

Partant de ces constats, nous avons décidé pour le projet de mutualiser nos connaissances et nos compétences, notamment en nous inspirant de l'existant, même s'il n'est pas encore très développé.
Pour ce qui est des corpus, nous allons nous appuyer sur les méthodologies employées pour les projets Picartext pour la langue picarde et BaTelÒc pour la langue occitane. Ces deux projets ont parallèlement débuté en 2006 (sans alors avoir connaissance l'un de l'autre) et visaient un même objectif : doter la langue picarde et la langue occitane d'une base de textes consultables en ligne.

La base de texte picarde, PicarText[1] (*Eloy et al*, 2015), a été réalisée au LESCLAP à Amiens sous la direction de Jean-Michel Eloy et Christophe Rey avec le soutien financier du conseil régional de Picardie. Elle comprend à l'heure actuelle environ dix millions de mots, de textes allant du XVIII^e^ au XXI^e^ siècle et de genres très variés (dictionnaires, contes, recueils de

1 https://www.u-picardie.fr/LESCLaP/PICARTEXT/Public/

poésie, romans, chansons). Les méthodes de recherche dans la base sont particulièrement intéressantes, car elles prennent en compte la variation graphique : on peut faire des recherches sous forme littérale ou avec des expressions régulières, mais il y a également des fonctionnalités qui intègrent la correspondance phonétique, ce qui permet de retrouver différentes formes orthographiques utilisées par les auteurs à condition que la prononciation soit identique. Il est également possible de prendre en compte la correspondance dialectale, c'est-à-dire retrouver les formes théoriquement possibles en picard, y compris avec d'autres prononciations que celle fournie. À cela s'ajoutent d'autres fonctionnalités de recherche : empan temporel, zone géographique, genre textuel.

La base BaTelÒc (Bras et Thomas, 2011) est réalisée au laboratoire CLLE-ERSS, Université de Toulouse 2 Jean Jaurès sous la direction de Myriam Bras. La première version de la base est en cours de finalisation. Elle contient environ trois millions de mots (85 œuvres d'une quarantaine d'auteurs), sur une période allant du XIX^e^ au XXI^e^ siècle, représentant des genres variés (contes, poésies, romans, nouvelles, mémoires) et relevant de plusieurs dialectes et de plusieurs graphies. À ce jour, tous les textes de la base ont été acquis au format numérique et encodés au format XML (TEI P5) avec le souci de rester le plus fidèle possible à l'édition papier (mise en forme). Les textes sont intégrés dans une base dotée d'une interface qui propose plusieurs outils de consultation. Le premier outil est une interface de sélection des textes qui permet de se constituer un corpus de travail (selon le titre, l'auteur, sa date de naissance, l'année de création ou d'édition de l'œuvre, le dialecte, la graphie, le genre…). Les autres outils sont des concordanciers qui permettent de rechercher les contextes d'emploi d'une forme ou de plusieurs formes avec des fonctionnalités telles que la forme « est », « contient », « commence par », « finit par » ou en exploitant le langage des expressions régulières.

2. Objectifs du projet RESTAURE

Le projet RESTAURE s'inscrit dans la complémentarité des projets PicarText et BaTelÒc et vise en premier lieu pour l'alsacien, l'occitan et le picard le développement de ressources et d'outils linguistiques. Il est prévu selon les états d'avancement pour chaque langue de constituer ou d'enrichir les corpus

de textes. La matière pour les trois langues concernées est abondante mais pas toujours disponible au format numérique. Un des objectifs visés par le projet est donc de numériser et *OCRiser*[1] une partie de cette matière. Nous souhaitons développer des outils *OCR* spécifiques à chaque langue et adaptés au traitement de la variation. Ces outils permettront de constituer et compléter nos corpus et seront également diffusés avec des licences libres à la fin du projet.

Le projet vise ensuite l'enrichissement de ces corpus avec des annotations linguistiques, en l'occurrence des annotations morphosyntaxiques. Cette annotation vise à associer à chaque forme des corpus, le lemme ou forme de citation (par exemple le verbe à l'infinitif, l'adjectif au masculin singulier, le nom au singulier), une catégorie grammaticale (nom, verbe, adjectif, adverbe…) et des informations morphosyntaxiques (personne, nombre, genre…). Ces annotations, dans le dispositif des bases de textes présentées ci-dessus, permettront de nouveaux modes de consultation des contextes d'emploi, par exemple la recherche de toutes les formes fléchies d'un verbe à partir de son lemme.

Nous avons fait le choix, dans le projet, d'utiliser des algorithmes par apprentissage, c'est-à-dire qui cherchent à apprendre des règles générales à partir d'exemples particuliers. Cela nécessite un travail pointu d'annotation des données. Dans le projet, les compétences linguistiques particulières à chaque langue régionale se trouvent dans les trois laboratoires qui développeront les annotations nécessaires. Néanmoins, nous souhaitons mettre en place un soutien commun sur toutes les compétences techniques et sur tous les aspects méthodologiques. Nous souhaitons nous doter d'une méthodologie commune qui pourra être également appliquée à d'autres langues qui souhaiteraient avoir le même parcours que le nôtre.

La recherche en traitement automatique des langues pour l'alsacien, l'occitan et le picard soulève des questions nouvelles sur le traitement de la variation qui sont incontournables pour doter chacune des langues d'une boîte à outils minimale en TAL. Nous œuvrons à la constitution de bases solides pour qu'ensuite de nombreuses applications puissent être développées (moteurs de recherche, correcteurs orthographiques, outils d'aide à la rédaction et à la traduction, synthèse vocale…).

1 *OCRiser* ou *océriser* : cf. supra, note 1, page 85.

Indications bibliographiques communiquées par les auteurs

Bernhard, D. (2014). Adding Dialectal Lexicalisations to Linked Open Data Resources: the Example of Alsatian, in *Proceedings of the Workshop on Collaboration and Computing for Under Resourced Languages in the Linked Open Data Era (CCURL 2014)*, May 2014, Reykjavík, Iceland. pp.23-29, 2014

Bernhard, D. et Ligozat, A.-L. (2013). Es esch fàscht wie Ditsch, oder net ? Étiquetage morphosyntaxique de l'alsacien en passant par l'allemand. In Actes de TALARE 2013 : Traitement Automatique des Langues Régionales de France et d'Europe, p. 209–220.

Bras, M. et Thomas, J. (2011). Batelòc : cap a una basa informatisada de tèxtes occitans. In A. Rieger (ed.) L'Occitanie invitée de l'Euregio. Liège 1981-Aix-la-Chapelle 2008 Bilan et perspectives, Actes du IX^{e} Congrès International de l'AIEO, Aache, Shaker.

Eloy, J.-M., Rey, C., Martin, F. (2015). PICARTEXT : Une ressource informatisée pour la langue picarde, Actes de *TALaRE 2015 - Traitement Automatique des Langues Régionales de France et d'Europe*, Juin 2015, Caen, France.

Vergez-Couret, M., Urieli, A. (2014). 'POS-tagging different varieties of Occitan with single-dialect ressources', Eds M. Zampieri, L. Tan, N. Ljubešić, J. Tiedemann, *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, Eds. (Dublin : Association for Computational Linguistics and Dublin City University), 21-29.