Communication sur les travaux de Òsca-Font dubèrta

Dominique Château-Annaud
Association Òsca-Font dubèrta,
1 charrèira/rue St Rames, 63000 Clarmont d'Auvèrnhe/Clermont-Ferrand
architecte logiciel et développeur
dc@macarel.net, d.chateau@laposte.net

Résumé. Cette communication présente l'intégration de deux développements informatiques récents conçus comme des outils linguistiques et lexicographiques séparés. Cette intégration se concrétise en un outil original, une plate-forme d'édition numérique dont la notion sera précisée.
Les deux projets sont implantés dans un site web et exploitent les données provenant de bases de données SQL. L'interface utilisateur est constitué de formulaires d'édition et de recherche, de tableaux en HTML et de rapports en différents formats. À l'origine les données proviennent de dictionnaires dialectaux, de listes de verbes, de modèles de conjugaison et d'autres informations annexes. L'ensemble est uniquement disponible dans un format faiblement structuré (traitement de texte WYSIWYG) impropre à un traitement numérique efficace, ce qui nécessite une conversion en base de données. Celle-ci a suscité beaucoup d'efforts et soulevé des contraintes méthodologiques et humaines.
Pour le conjugueur automatique les algorithmes sont codés comme une hiérarchie de classes d'objets facile à adapter pour d'autres dialectes [1] et extensible à d'autres formats de sortie.
Pour conclure nous évoquerons l'extension des capacités de la plate-forme vers les bases de données textuelles NoSQL et vers une architecture REST.

Abstract.

Paper on the work of OSCA-Font dubèrta
This paper presents the integration of two recent IT developments designed to be two separated linguistic and lexicographical tools in an original one, a digital publishing platform, which concept will be described.
Both projects are web applications, typically LAMP (Linux, Apache, MySql, PHP). The first one is designed to build a transdialectal lexical base. Data comes from dialectal dictionaries, verb lists, conjugation patterns and other related information to be converted in database.
Despite the weakly structured format (WYSIWYG word processing) not usable for serious digital processing, the conversion populated a big lexical base. The conversion task drew a lot of efforts and raised methodological and human constraints.
The second one is an automatic conjugator made from an easy to adapt object-oriented hierarchy for the lengadocian occitan dialect. One more dialect, gascon is available and wait to be tested. Provencal dialect is on study. Output formats extensions can be implemented by a loosely coupled coding.
As a conclusion we will discuss the extension of the publishing platform capabilities geared to textual NoSql databases and REST architecture.

Mots-clés : plate-forme d'édition numérique, base de données NoSQL, outil lexicographique, conjugueur automatique, conception orientée-objet, lexique, dictionnaire.

Keywords: digital publishing platform, NoSQL database, lexicographical tool, automatic conjugator, object-oriented design, glossary, dictionary.

1. le gascon est en test, le provençal en étude

Introduction

La langue occitane compte six dialectes, d'ouest en est, le limousin, l'auvergnat, le vivaro-alpin pour la moitié septentrionale ; le gascon, le languedocien et le provençal couvrent la partie méridionale. Des isoglosses traversent le domaine occitan et séparent assez précisément le sud du nord. Ainsi le traitement du C + A latin et du G + A latin produisent respectivement pour le sud et pour le nord :

— canta / chanta
— galina / jalina

Dans la partie médidonale entre autres, les isoglosses du F / H latin séparent l'ensemble dialectal languedocien du gascon :

— filha / hilha

plus à l'est la confusion V / B caractérise le dialecte languedocien et se distingue du provençal qui différencie les deux vocalisations (Dupuy, 1972).

Dans la partie septentrionale, les distinctions sont plus ténues entre le limousin et l'auvergnat puis entre l'auvergnat et le vivaro-alpin, nous ne les évoquerons pas ici.

La présentation à gros trait faite plus haut peine à traduire la richesse de ces variétés dialectales. Il est impératif de les révéler et de les comparer en mettant en évidence leur tronc commun preuve de l'intégrité linguistique de l'ensemble. Malgré leur dissémination, des dictionnaires ou des lexiques spécifiques à ces dialectes ne manquent pas pour servir de référence. Cependant nous constatons une grande hétérogénéité méthodologique et des moyens très inégaux, ce qui affecte la représentativité de certains dialectes (tableau 1). La présentation des dictionnaires doit dépasser ces différences et doit s'efforcer de proposer un traitement équitable de toutes les variétés dialectales.

dialecte	**nombre d'entrées**	**pourcentage**
languedocien	34 002	23,95 %
gascon	63 513	46,14 %
vivaro-alpin	23 235	16,36 %
provençal	12 012	8,46 %
auvergnat	7 218	5,08 %
Total	141 980	100 %

TABLE 1 – Représentation des dialectes par leur nombre d'entrées

Plate-forme d'édition numérique Le projet présenté est la fusion de deux développements séparés conçus pour être des outils linguistiques et lexicologiques : un lexique et un conjugueur automatique. De cette intégration résulte une plate-forme d'édition numérique, c'est cette idée originale qui est décrite par l'article.
Préalablement au premier développement, une base lexicale est constituée par la conversion de plusieurs dictionnaires dialectaux provenant des cinq dialectes sur les six qui composent l'occitan. Incorporées dans une base de données, ces données lexicales sont interrogeables par des requêtes SQL. Elles sont consultables par le public mais elle servent également pour la création des lexiques interdialectaux Basicòt/Basic. Ce dernier usage est réservé à des spécialistes disposant de droits d'accès particuliers aux données, afin d'aider à la construction de ces lexiques un éditeur interactif permet aux lexicographes de créer, supprimer et modifier des entrées. L'interface de l'éditeur recherche les occurrences du terme à traiter dans la base lexicale, les résultats retournés aident à la décision des lexicographes.
Le conjugueur automatique enrichit la base lexicale par son exécution, et en retour le conjugueur peut fouiller la base lexicale pour afficher les définitions et les traductions éventuelles. Dans l'état actuel du développement plusieurs conjugueurs sont en chantier :

— Lengadocian (Sauzet, 2015) en production, nouvelle version en test ;
— Gascon (Bianchi & Viaut, 1995) en développement ;
— Provençal (Moulin, 2005) à l'étude.

La plate-forme intègre les outils en un seul site web et présente leurs résultats en différents formats, la variété de formats peut convenir au grand public, comme au lexicographe soucieux de vérifier la sortie imprimée de son lexique, d'autres sites web ou d'autres applications à venir (sur tablettes ou téléphones intelligents) questionneront la base lexicale au moyen d'URL [2]. La variété des formats de sortie garantit la polyvalence de la plate-forme d'édition numérique et constitue son originalité.

Pour conclure nous dresserons un état actuel du développement et énoncerons les orientations et les pistes de développement qui se dessinent.

Nota : Distinction entre dictionnaires et lexiques, définitions Dans cet article nous ne faisons pas la différence entre lexiques et dictionnaires. Souvent les auteurs [3] qualifient un peu abusivement leur ouvrage, de dictionnaire alors qu'ils ne comportent pas d'exemple, de mise en contexte ou de définition détaillée. Nous emploierons le terme :

— « dictionnaire » indifféremment dans tous les cas sauf pour le Basicòt/Basic qualifié de lexique ;
— « définition » pour la partie restante d'une entrée dont on a isolé le terme et la catégorie grammaticale ;
— « base lexicale » pour l'ensemble des dictionnaires stockés en base de données.

1 Conversion des dictionnaires dialectaux et outil de création de lexique

1.1 Génèse du projet

Le besoin s'est fait sentir de rassembler les dictionnaires qui témoignaient de la richesse dialectale de notre langue et d'utiliser les nouvelles technologies de l'information afin de les proposer au public dans un site de référence.
Un événement a aiguillonné ce besoin lorsque est apparu sur la toile, un dictionnaire fantaisiste [4] promouvant un occitan de communication peu respectueux des variétés dialectales. Ce dictionnaire en ligne instillait une utilisation pernicieuse des outils informatiques au détriment d'une langue menacée. Cet épisode a fait prendre conscience de la précarité d'une langue minorisée devant une technologie puissante qui peut la desservir et la pervertir si l'on ne se donne pas la volonté de l'utiliser avec détermination avant que d'autres ne la dévoient.

La communauté occitane s'est mobilisée et plusieurs projets ont relevé le défi, mentionnons l'**Academia Occitana-Consistòri del Gai Saber** [5] qui propose une réponse similaire à celle du Congrès. Le Congrès Permanent de la Langue Occitane est un organisme de régulation de la langue occitane, son site web est la vitrine de ses travaux et il héberge les dictionnaires numérisés.
La première tache a été de mettre en ligne avec diligence les dictionnaires que les auteurs, leurs ayants droit et les éditeurs acceptaient de nous confier [6]. La priorité a été mise sur les dictionnaires Français-Occitan afin d'entamer ensuite le développement du Basic/Basicòt.
Depuis peu, tous les dictionnaires Français-occitan convertis sont publics, ainsi cinq des six dialectes sont accessibles en ligne. Les utilisateurs des dictionnaires peuvent interroger les termes français et voir la traduction correspondante dans le dictionnaire de leur choix, selon le dialecte de leur choix. Il est toujours possible d'élargir la recherche aux autres dictionnaires s'il n'y a pas de résultat pour la sélection demandée. Une recherche dans les définitions occitanes est également possible (recherche plein texte). Le CPLO a poursuivi le travail d'incorporation des dictionnaires *Occitan-Francés* dans la base lexicale qui en compte cinq à l'heure actuelle mais nous resterons sur les dictionnaires Français-Occitans. L'application web de recherche dans la base lexicale du CPLO est nommée « Dico d'Òc » [7].

2. Uniform Resource Locator
3. Dans ces lignes les agents auteurs, opérateur, utilisateurs, etc s'entendent au féminin comme au masculin
4. PanOccitan.org, l'occitan de communication
5. http ://www.academiaoccitana.eu
6. Saluons le travail opiniâtre de Gilbert Mercadier (président du CPLO) en qualité de négociateur
7. http ://locongres.org/fr/applications/dicodoc-fr/dicodoc-recherche

Nom du dictionnaire	code	nombre d'entrées
Laus (languedocien)	LAUS (Laus, 2005)	34 002
Rei Bèthvéder (gascon toulousain)	RBVD (Rei-Bèthvéder, 2004)	13 998
Atau que's ditz ! (gascon)	ATAU (ouvrage collectif, 1998)	7 303
Per Noste (gascon)	PNST (Miquèu Grosclaude & Guilhemjoan, 2007)	44 212
Faure (vivaro-alpin)	ALPC (Faure, 2009)	23 235
CREO Provença (provençal)	PROV (Elie Lèbre & Moulin, 1992)	12 012
Omelhier (Auvergnat)	OMLH (Omelhièr, 2004)	7 218
Total		141 980

TABLE 2 – Liste des dictionnaires convertis Français-Occitan et leur nombre d'entrées

1.2 Le procédé de conversion

Les documents sont originellement écrits à l'aide d'un traitement de texte à l'exception du dictionnaire provençal qui est écrit en LaTeX. Le procédé est décomposé en phases :

1. Les documents issus d'un traitement de texte sont convertis en fichier HTML.
2. Le fichier HTML est traité par un script Perl qui utilise une série d'expressions rationnelles pour corriger, épurer, formater les entrées du dictionnaire en enregistrements sur une seule ligne. Le fichier résultant est constitué de lignes d'entrées de dictionnaire, les éléments d'information sont balisés par un langage de marquage inspiré [8] de Docbook (dialecte de XML). Ce processus de moulinage est raffiné itérativement jusqu'à l'obtention d'un fichier exempt de marques HTML et d'entrées de dictionnaire incomplètes, c'est le fichier transitoire (voir fichier transitoire figure 1).
3. Le fichier transitoire est traité par un script Perl qui vérifie l'intégrité de la structure des entrées du dictionnaire (entre les deux balises `<glossentry></glossentry>`). Cette structure est élémentaire, une expression rationnelle suffit pour en tester la cohérence. Un fichier d'erreur est généré.
4. Un script Perl interprète les entrées `<glossentry></glossentry>` afin de les convertir en énoncés SQL d'insertion dans la base de données.

La validation et la génération SQL (phases 3 et 4) peuvent être implantées dans un seul script.
L'interprétation du fichier d'erreur donne des indications permettant d'amender le script de la phase 2 avec éventuellement la mise en place d'un dispositif de correction d'erreurs.

1.3 Justification du procédé

La conversion Word vers HTML procède au changement des balises internes au traitement de texte en balises HTML distingués par des attributs de style. Ces attributs sont nécessaires au rendu de la mise en forme similaire à celui du document original. Les attributs qui nous intéressent, que nous appellerons les marqueurs sont ceux qui séparent en les encadrant, trois éléments d'information fondamentaux, constitutifs d'une entrée de dictionnaire :

— le terme (ou la vedette)
— la catégorie grammaticale relative au terme
— La définition

Le traitement de texte utilisé pour créer les documents sources est presque toujours Microsoft® Word. Donc c'est ce logiciel qui sera employé pour faire la première conversion en HTML, plusieurs options de conversion sont proposées, on choisira celle de plus bas niveau ayant le moins de styles. OpenOffice peut également servir d'outil de conversion, cependant il donne un résultat peu exploitable à cause de la multiplication des styles rendant la phase 2 inapplicable. La conversion document word-HTML avec Microsoft® Word 2003 est la solution retenue.
Il faut savoir que les marqueurs qui distinguent les éléments d'information ne sont pas constants dans le fichier HTML

8. Docbook dispose de balises pour définir une entrée de lexique mais son vocabulaire est très insuffisant pour nos besoins. Son utilisation a été envisagée mais non retenu à cause du peu d'engouement qu'il suscite en dehors des lecteurs de la documentation de FreeBSD.

```
...
<glossentry><glossterm>abaissement</glossterm><genregram>nm</genregram><glossdef>abaissamen
(nm), diminucion (nf), demenia (nf).</glossdef></glossentry>
<glossentry><glossterm>abaisser</glossterm><genregram>v</genregram><glossdef>abaissar,
abeissar, baissar, clinar, demenir.</glossdef></glossentry>
<glossentry><glossterm>abandon</glossterm><genregram>nm</genregram><glossdef>abandon
(nm), renonciacion (nf).</glossdef></glossentry>
<glossentry><glossterm>abandon (à l')</glossterm><genregram>loc
adv</genregram><glossdef>a la picorea (F).</glossdef></glossentry>
<glossentry><glossterm>abandonner</glossterm><genregram>v</genregram><glossdef>abandonar,
renonciar, laissar, quitar. "Aqueu pichon s'abandona" : se dich d'un
pichon que, per lo prumier còp, fai quauques passes sensa estre
sostengut.</glossdef></glossentry>
<glossentry><glossterm>abasourdir</glossterm><genregram>v</genregram><glossdef>esbalordir,
estabosir, espantar, encornorir, encocornir.</glossdef></glossentry>
...
```

FIGURE 1 – Extrait de fichier transitoire

(voir Critique du traitement de texte, WYSIWYG contre WYSIWYM §1.3), c'est la raison de la mise au point itérative de la phase 2 à partir du fichier d'erreur produit à la phase 3.
Dans certains rares cas il peut y avoir collision dans les marqueurs. La conséquence en est la production d'entrées mal formées, mais plus fâcheusement, des entrées tronquées qui s'étendent sur plusieurs lignes. La correction humaine est alors requise lors d'une relecture.

Critique du traitement de texte, WYSIWYG contre WYSIWYM

Définition de WYSIWYG *What You See Is What You Get*, cet acronyme anglo-saxon désigne les traitements de texte à usage domestique pour lesquels la mise en forme est confondue avec le contenu.

Définition de WYSIWYM *What You See Is What You Mean*, cet acronyme anglo-saxon désigne les chaînes éditoriales à la LaTeX.

Séparation du fond et de la forme La confusion entre le fond et la forme propre aux traitements de texte type WYSIWYG est le nœud des difficultés que nous rencontrons dans la conversion. La critique du WYSIWYG est connue, les auteurs sont distraits par la mise en page tandis qu'ils devraient rester concentrés sur la structure et le contenu du document. Mais outre cette critique d'autres problèmes se font jour dans le cadre précis de l'écriture d'un document contenant des données structurées et répétitives, notamment trois problèmes.

1. L'utilisation peu rigoureuse de la mise en forme occulte la nécessité de qualifier les éléments d'information correctement. On se contente de graisser, de mettre en italique ou d'utiliser une police de caractères particulière pour distinguer le terme, la catégorie grammaticale ou les phrases de contexte incises dans la définition. D'autres éléments d'information secondaires restent indifférenciés (exemple : 1)
2. La mise en forme s'effectue sur la sélection à la souris avec le risque d'incorporer des ponctuations entre les balises et polluer ainsi l'élément d'information. Pire, on retrouve des cas où l'auteur sélectionne des éléments d'information mitoyens pour y appliquer une mise en forme amalgamant les deux, polluant l'un et perdant l'autre (exemple : 2).
3. L'insertion de fichiers traités à part avec une autre configuration de traitement de texte peut entraîner la génération de styles insidieusement différents dans le fichier source. Cela a pour effet d'affecter les marqueurs servant à la pose des balises.
   Nous avons eu récemment l'expérience d'une lettre entière qui avait échappée au script de conversion car les marqueurs ayant changés, ils n'ont pas été détectés par les expressions rationnelles. Cela provient du fait que cette lettre avait été traitée à part (dans un autre fichier) puis incorporée tardivement à l'ensemble.

exemple :

1. `<i>f, </i>` pour `<i>f</i>,` . La mise en italique englobe la virgule et son espace.
2. `<b>II pendent que</b>` au lieu de `<b>II</b> <b>pendent que</b>`. La graisse est appliquée indistinctement à la numérotation et à la locution de conjonction occitane. Cependant à l'affichage à l'écran le résultat est le même. L'auteur n'a pas conscience de sa bévue.

Contraintes Il est évident que pour l'évolution du dictionnaire, le traitement de texte n'est plus viable dès lors que les données sont reportées dans une base de données car il n'est pas raisonnable de repartir dans une itération de correction en traitement de texte, de conversion, d'incorporation dans la base. Nous recommandons d'abandonner cet outil et de saisir les ajouts ou les modifications au moyen d'un formulaire de saisie rigoureux disposant de mécanismes de vérification et d'auto-complétion. Ces mécanismes sont à élaborer en fonction de la méthodologie convenue par les lexicographes et avec eux. Un formulaire accessible en ligne présente l'avantage de devenir un outil collaboratif dès lors qu'on peut gérer des accès contrôlés en modification. Avec des cadres applicatifs (*frameworks*) sophistiqués JOOMLA, DRUPAL, des accès avec des droits différents sont possibles, un forum et un système de vote également.
Un problème humain se pose néanmoins celui de certains lexicographes qui répugnent à se servir de l'outil informatique aussi ergonomique soit-il, il n'est pas rare que ces derniers confient leurs travaux à des opérateurs chargés de les taper. Une médiation est donc nécessaire et un soin tout particulier doit être mis au confort de l'opérateur/correcteur/lexicologue qui fera la transcription. En retour il convient de faire parvenir aux lexicologues non informatisés les résultats des corrections dans un format le plus proche possible du résultat final. Dans notre cas, il s'agit d'une édition PDF tirée d'une mise en forme LaTeX du dictionnaire prêt à imprimer. La génération de ce PDF est semi-automatique dans l'état actuel du développement et réclame l'intervention de l'administrateur, mais pourra bientôt être lancée à loisir par l'opérateur de saisie.

1.4 Utilisation de la base lexicale dans la construction du Basicòt/Basic

Dès que la base lexicale composée de tous les dictionnaires dialectaux est en place, elle peut servir de référence à la constitution du Basicòt. Le Basicòt est un lexique interdialectal [9] proposant une norme conformément à la mission de régulation linguistique du CPLO [10]. Le Basicòt porte sur deux dialectes, gascon et languedocien. L'outil de construction du Basicòt affiche les définitions provenant de la base lexicale pour chaque terme français à traiter.

Basic Le Basic suit le même principe mais on procédera pour les six dialectes de l'espace occitan. Dans l'état actuel de l'organisation c'est une tâche trop intense pour les lexicographes, elle a été différée.

1.4.1 Principe

Un liste de termes provenant du Petit Robert Élémentaire constitue un fond d'entrées françaises au lexique en création, la base de données compte au départ près de 12 000 entrées à compléter. Le site web hébergeant la plate-forme d'édition numérique propose un formulaire pour parcourir ces entrées, voici sa description.
L'opérateur accède au terme désiré soit par la boîte de recherche, par les boutons de navigation, par le tableau extrait (voir extraction lettre par lettre §1.4.2). Pour chaque terme une recherche dans la base lexicale retourne les entrées provenant de chacun des dictionnaires dialectaux si la recherche est fructueuse. On se trouve dans la même situation qu'un utilisateur du Dico d'Òc. Le lexicographe peut alors consulter les dictionnaires gascons et languedocien d'un seul coup d'œil car le Basicòt se limite à ces deux dialectes. Il est possible de créer des acceptions, des homographes. Un mécanisme de validation par un superviseur est possible, permettant de basculer un booléen lorsque le terme est terminé et stable. Ce booléen « *acabat* » (terminé) doit être vrai pour permettre à la suppression d'une entrée, ceci par mesure de sécurité.
Pour chaque entrée, les règles suivantes s'appliquent :

— si une forme est utilisée dans tous les domaines dialectaux elle est qualifiée de *globalement* usitée ;

9. Basicòt/Basic sont considérés comme des lexiques car leurs définitions ne contiennent que les termes traduits.
10. http ://locongres.org/fr/lo-congres-fr/l-institution/missions

— pour le Basicòt, si une forme est utilisée dans les deux domaines dialectaux elle est qualifiée de *communément* usitée pour LG ;
— si un dialecte ne connaît pas la forme usitée dans le reste de l'espace, celle-ci est qualifiée de *communément* usitée ;
— dans tous les cas les formes vernaculaires sont saisies en mentionnant leur provenance.

La base de termes français est enrichie graduellement tandis que se raffine la méthodologie lexicographique des contributeurs.

Sur la fiche d'une entrée, il est possible de saisir les informations, mais on peut aussi ajouter une acception si nécessaire (le genre grammatical demeure le même). Si c'est un homographe, une nouvelle entrée doit être créée et une nouvelle catégorie grammaticale doit être choisie. Ensuite le lexicographe entre le terme commun aux deux dialectes dans le champ « Commun LG ». Le champ « Sorga (*source*) » affiche le résultat de la recherche dans la base lexicale. Le champ « Còde de sasida (*code*) » est utilisé pour saisir les formes vernaculaires si elles existent. Les termes sont saisis entre les deux balises XML correspondant au dialecte. Un script javascript intercepte la frappe et affiche le terme dans la couleur spécifique du dialecte. Le lexicographe vérifie la bonne saisie en constatant le changement de couleur dans l'affichage de la définition dans la boîte à droite.
La chaîne se presente ainsi :

```
<mjrn>
     <leng></leng>
     <gasc></gasc>
     <prvc></prvc>
</mjrn>
<sept>
     <auv></auv>
     <valp></valp>
     <lim></lim>
</sept>
```

C'est cette chaîne XML qui est enregistrée. `<mjrn></mjrn>` signifie méridional (*miègjornal*) `<sept></sept>` signifie septentrional, ces rubriques regroupent les dialectes des deux zones.

1.4.2 Extraction des données

L'extraction des données est la partie publication de la plate-forme. C'est cette fonctionnalité qui offre la restitution des données de la base lexicale garantissant un accès large au public et aux chercheurs dans le maximum de formats demandés. Pour les opérateurs c'est un autre moyen d'accéder aux données pour afficher celles-ci par lettre et sous forme de tableau, la flèche verte est un bouton permettant d'ouvrir le formulaire de modification pour le terme courant. Ainsi le lexicographe peut parcourir le document en entier et faire les modifications à la volée. Il peut à loisir copier-coller ces informations et obtenir des données statistiques lui permettant de vérifier l'avancement du projet. Nous n'en présentons l'extrait d'un seul format (HTML+css).

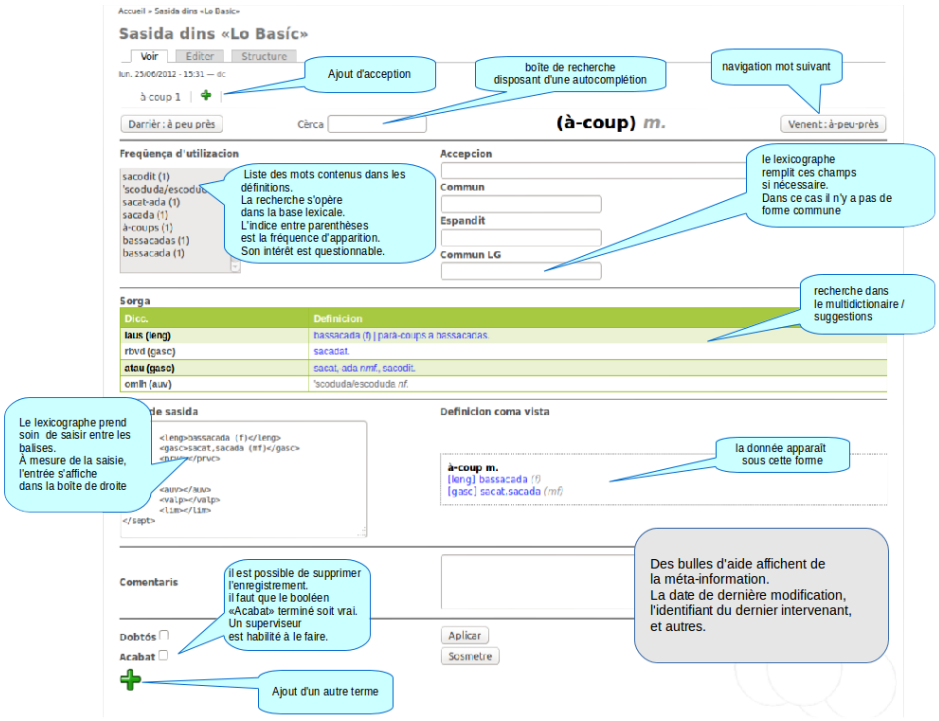

FIGURE 2 – Interface de saisie pour les lexiques (copie d'écran)

2 Le conjugueur automatique

Le projet de conjugueur languedocien est une initiative du **CPLO**. Patrick Sauzet met à notre disposition une liste de verbes considérable, chaque verbe est associé à un modèle. Cette catégorisation est très riche, le nombre important de modèles témoigne de l'acuité de la recherche de Sauzet.
Notons que ce travail est un projet original différent du travail précédent (Sauzet & Ubaud, 1995), à l'heure où nous écrivons ces lignes l'ouvrage est en cours de correction et en mise en forme, sa parution est prévue dans l'année. L'ouvrage de Patrice Poujade a été également consulté (Poujade, 2005) pour certaines vérifications.

2.1 Quatre algorithmes pour les trois groupes + auxiliaire

L'analyse de ce projet fait ressortir l'asymétrie dans le nombre de verbes (la cardinalité) répartis sur les quatre groupes. Le degré de complexité de conjugaison est paradoxalement moins élevé pour les verbes du deuxième groupe que ceux

Groupe	Nombre de modèles	Nombre de verbes	Verbes/modèle
auxiliaire	5	18	3,60
1er groupe	47	10 743	228,50
2e groupe	8	1 299	162,50
3e groupe	70	476	7,00

TABLE 3 – Cardinalités des verbes et répartition des modèles

du premier. Le premier groupe en languedocien est très fouillé, il se caractérise par nombre de modèles décrivant les alternances vocalique affectant la racine. Tandis que le deuxième groupe est le plus parcimonieux en modèles.
Les auxilliaires utilisent un modèle pour moins de quatre verbes. Le troisième groupe est tout aussi dispendieux.

FIGURE 3 – Une partie de l'affichage colorisé

Différences entre les algorithmes des premier et deuxième groupe avec le troisième et auxiliaire Les algorithmes du premier et second groupe contiennent toutes les désinences dans quelques tableaux sans apport extérieur. L'intégralité de la logique de la conjugaison des verbes du premier et second groupe est contenue dans le code de l'algorithme.
Ce n'est pas le cas pour les algorithmes des auxiliaires et du troisième groupe. Les désinences irrégulières sont tirées de la base de données. Ainsi l'algorithme est relativement plus simple mais sa généricité est décevante. On peut qualifier ce traitement de « mode dégradé ».
Malgré les efforts pour obtenir des algorithmes le plus génériques possibles, demeurent des traitements dépendants de l'identifiant du modèle. Ainsi les numéros de modèles, codés dans le programme, ne peuvent évoluer au risque d'entraîner un réécriture des algorithmes.

2.2 Fonctionnement des algorithmes

La structure générale des algorithmes est la suivante :
Selon le modèle du verbe, on détermine le groupe de celui-ci et on invoque l'algorithme spécifique de ce groupe, on procède à la racinisation et on accole mode par mode, temps par temps, personne par personne la désinence à la racine. Un traitement particulier doit être pratiqué sur la racine selon les besoins. Par exemple l'alternance *e/è* en languedocien ne se produit que pour certains temps et certaines personnes, il a une régularité dans les cas particuliers. Considérons cette règle.

Temps	Personnes
Présent de l'indicatif Présent du subjonctif	1 S
	2 S
	3 S
	3 P
Impératif (forme affirmative) Impératif (forme négative) == présent du subjonctif	2 S

TABLE 4 – Application de l'alternance vocalique

Dans le cas d'une alternance *e/è*, le dernier « e » du radical deviendra « è » pour les temps et les personnes du tableau.
Également les règles phonologiques s'appliqueront systématiquement comme pour l'alternance *g/gu*.
Citons pour exemple le modèle 114 : *conjugason 1 alternanta è/e, g/gu* dont le verbe type est **negar** [noyer (se)]. Le trai-

tement sur le radical s'effectuera comme décrit ci-dessus pour l'alternance *è-e*. Le traitement phonologique s'appliquera sur la totalité de la conjugaison pour l'alternance *g/gu*.

Érosion du radical Les verbes irréguliers du troisième groupe sont également racinisés mais le radical n'est pas stable, il est particulièrement érodé dans certains cas pour la troisième personne du présent de l'indicatif, de la deuxième personne de l'impératif pour ne citer qu'eux. La logique trop complexe à implanter dans le code est suppléée par les désinences provenant de la base de données.
Pour creuser la racine nous disposons d'un caractère spécial affixé à la désinence qui provoque un effacement arrière permettant ainsi d'éroder le radical et dans le pire des cas d'y substituer le reste de la racine et la désinence.
En guise d'exemple extrême, le verbe **fúger** est très irrégulier. Il illustre le caractère peu glorieux de cette solution. En effet seule l'accentuation du « *ú* » demeure à l'infinitif, il n'apparaît plus dans le reste de la conjugaison qui est entièrement recomposée à partir de la base de donnée, c'est le choix de prioriser la racinisation, dans ce cas seul l'infinitif racinisé est irrégulier. On conserve ce verbe pour mémoire, et c'est une occasion pour proposer une alternative sous la forme d'un « renvoi » (voir 2.3 Le renvoi) vers le verbe **fugir**.

2.3 Le renvoi

Le renvoi est un bouton ou un hyperlien que apparaît dans l'affichage de la fiche d'une conjugaison s'il s'agit d'un francisme comme dans le cas de *acochar* pour accoucher alors que le verbe recommandé est *ajaire*, dans le cas d'un usage impropre **difusir* au lieu de *difusar*. Quand il clique sur le bouton l'utilisateur est renvoyé vers le verbe recommandé. Le renvoi est univoque, il n'y a pas de retour à la page précédente, il fait figure de préconisation, sous forme d'incitation bienveillante. Un bouton ou un hyperlien est une invitation non contraignante à visiter la fiche du verbe recommandé.

2.4 Les verbes frères et autres fonctionnalités

Lorsqu'un verbe appartenant à un modèle non pléthorique est affiché à l'écran, les autres verbes appartenant au modèle sont également accessibles, ce sont en principe les verbes du troisième groupe qui présentent un grand éventail de modèles et un faible ratio verbes/modèle qui bénéficient de cette fonctionnalité.
On affichera un tableau des verbes courants et un palmarès des verbes le plus demandés à la consultation, il est possible que ces statistiques aient un quelconque intérêt pour les lexicographes.
Certains sites de conjugaison ont la bonne initiative de proposer des exercices pour pratiquer les conjugaisons. L'intérêt pédagogique est évident, c'est une voie à suivre pour de prochaines fonctionnalités.

2.5 Validation et correction du conjugueur languedocien

La phase de correction et de validation a été assurée par Bernard Moulin et Florence Malcouyre. Le cycle de validation s'est mis en place autour du site web de développement, avec un rapport de bogues mis à jour à chaque correction faisant office de *release notes* et affiché en préambule. Pour la correction des verbes du troisième groupe, une conjugaison exhaustive a été lancée et capturée dans un fichier CSV exploitable dans une feuille de calcul rendant aisé la visualisation des problèmes touchant soit l'algorithme, soit les données dans la base de données. Une fois les corrections des algorithmes et des données ayant été faites, les problèmes d'ordre linguistique et méthodologique [11] ont été référés au conseil linguistique du CPLO par Mr Bernard Moulin.

2.6 Les conjugueurs des autres dialectes

L'algorithme du conjugueurs gascon (Bianchi & Viaut, 1995) est implanté et en cours de test. Celui du provençal (Moulin, 2005) est à l'étude, grâce à la modularité du code, l'implantation n'est pas longue mais il appartient à une équipe d'experts de catégoriser la liste de verbes selon les modèles définis par P. Sauzet, c'est une opération laborieuse nécessitant plusieurs validations.

11. conjuguaison des verbes impersonnels

	Bianchi+Astié	Sauzet
grop 1	14 438	10 743
grop 2	2 000	1 299
grop 3	798	476
Total	17 236	12 538

TABLE 5 – Nombre des verbes listés

2.7 Intégration des conjugueurs dans la plate-forme

2.7.1 Arrimage aux dictionnaires dialectaux

La présentation actuelle du conjugueur pêche par le manque de mise en contexte des verbes conjugués. Si la signification du verbe *Cantar* ne cause pas de problème pour un occitanophone, on peut se poser la question pour *decopar 1* (Modèle N°120 conjugaison alternante ò/o) et *decopar 2* (Modèle N°100 conjugaison non alternante). Pour *decopar 1* un verbe alternatif est *talhar* (tailler) mais quelle est la signification et l'acception de *decopar 2*. Une recherche dans la base des dictionnaires dialectaux serait la bienvenue pour proposer l'acception, la définition voire la traduction du verbe courant.

2.7.2 Publication

La hiérarchie de classes PHP dispose à présent d'un autre objet de la hiérarchie de classe responsable de créer une sortie LaTeX qui est ensuite compilée à la volée pour créer la fiche du verbe conjugué dans un fichier PDF. L'utilisateur peut télécharger ce fichier à sa guise, l'imprimer, le partager, l'envoyer par courriel. Cette fonctionnalité utilise le concept de chaîne éditoriale comme nous l'avons vu. Dans le cas présent la génération de PDF issu de LaTeXest automatique, si le fichier PDF n'existe pas pour le verbe, le fichier est généré ainsi la banque de fichiers PDF croît à mesure de l'usage.

3 Conclusion

Le développement de la plate-forme de publication numérique a passé l'étape de la preuve de concept avec la génération des PDF mise en forme par LaTeX. Pour le Basicòt, l'opérateur pourra bientôt lancer la production d'une épreuve du lexique à sa guise. Les conjugueurs ont déjà passé ce cap, même si la génération de fichier LaTeX est moins complexe, les fiches sont générées automatiquement.

Dès la preuve de concept réalisée, l'orientation à prendre est claire et les nouvelles fonctionnalités découlent naturellement :

— L'exploitation des données doit continuer en reliant les dictionnaires entre eux par une clef unique comportant le terme, sa catégorie grammaticale uniformisée et sa source. L'utilisation du *Part of Speech* s'impose.
— À partir de la base lexicale et des conjugueurs, une indexation massive permettra d'offrir des possibilités de recherche inédites bien supérieure aux fouilles SQL, on parle ici de bases de données textuelles NoSQL (*Not Only SQL*) (Grainger & Potter, 2014) (Grant S. Ingersoll & Farris, 2013).
— Parallèlement, la plate-forme doit exporter ses résultats de façon structurée et uniforme, certes les résultats sont affichés à l'écran sous forme de rapports divers et variés mais il est également intéressant de communiquer avec elle au moyen de requêtes URL, base de l'architecture REST (*representational state transfer*). Ce mode de communication est également connu comme REST APIs, le format de sortie est un objet JSON, l'ensemble des clauses des URL avec leurs attributs constitue l'API (*Application Programming Interface* / interface de programmation). Il est envisageable de regrouper les services rendus par la plate-forme dans cette API. Citons que de grands acteurs des technologies de l'information offrent de telles APIs.
— Un format de sortie de choix à explorer est le TEI (*Text Encoding Initiative*) destiné à enrichir la polyvalence de la plate-forme.

Remerciements/Mercejaments

Nous tenons à remercier Gilbert Mercadier, président du CPLO ;
Patrick Sauzet professeur de linguistique occitane à l'université Jean Jaurès de Toulouse et président du conseil linguistique du CPLO ;
Florence Malcouyre, lexicographe au CPLO, correctrice du *Basicòt* et des conjugueurs automatiques ainsi qu'à tous les correcteurs impliqués dans le Basicòt.

Références

BIANCHI A. & VIAUT A. (1995). *Fichas de grammatica d'occitan gascon normat t1, prononciacion e grafia conjugasons*. 33405 Talence CEDEX : Presses Universitaires de Bordeaux.

DUPUY A. (1972). *La petite encyclopédie occitane*. SABER, Montpellier.

ELIE LÈBRE G. M. & MOULIN B. (1992). *Dictionnaire de base français-provençal*. CREO Provença.

FAURE A. (2009). *Diccionari Alpin d'Òc (vivaroaupenc)*.

GRAINGER T. & POTTER T. (2014). *Solr in Action*. Manning Publishing Co.

GRANT S. INGERSOLL T. S. M. & FARRIS A. L. (2013). *Taming Text How to Find, Organize, and Manipulate It*. Manning Publishing Co.

LAUS C. (2005). *Dictionnaire Français / Occitan (Languedocien)*. Castres : IEO del Tarn.

MIQUÈU GROSCLAUDE G. N. & GUILHEMJOAN P. (2007). *Dictionnaire Français / Occitan (Gascon)*. Per Noste edicions.

MOULIN B. (2005). *Grammaire occitane, le parler bas-vivarois de la région d'Aubenas*. IEO section vivaroise.

OMELHIÈR C. (2004). *Petit dictionnaire français-occitan d'Auvergne*. Ostal del libre - Collection Parlem.

OUVRAGE COLLECTIF (1998). *Atau que's ditz*. association PARLEM.

POUJADE P. (2005). *Los vèrbs conjugats, Memento verbal de l'occitan*. 09100 Pamiers : IEO Arièja.

REI-BÈTHVÉDER N. (2004). *Dictionnaire Français / Occitan Gascon Toulousain*. Toulouse : IEO edicions.

SAUZET P. & UBAUD J. (1995). *Le verbe occitan / Lo vèrb occitan guide complet de conjugaison selon les parlers languedociens*. Aix en Provençe : EDISUD.

FIGURE 4 – *Taton lo mascòt d'Òsca*