TALaRE 2015

Table des matières

Création de ressources lexicales pour une langue d'oïl : le parlanjhe

Marie-Hélène Lay[1,2], Jean-Christophe Dourdet[1,2,]
(1) Université de Poitiers, 1, Rue Raymond Cantel Bât A3, TSA 11102, 86073 POITIERS CEDEX 9
(2) Laboratoire FoReLL, EA 3816
marie-helene.lay@univ-poitiers.fr, jean-christophe.dourdet01@univ-poitiers.fr

Résumé.

Le présent article porte sur la constitution de ressources lexicales pour le poitevin-saintongeais, langue régionale (très) faiblement dotée. Depuis 2006, le projet TelPoS (Textes Electroniques en poitevin-saintongeais) a permis la constitution d'une base de données de textes (essentiellement littéraires) caractérisés par une forte variation, tant diatopique que diachronique (le premier texte date du 16e siècle). Le parlanjhe est une langue d'oïl, donc morphologiquement proche du français : nous avons choisi d'adapter des ressources dont nous disposions pour le français, en intégrant à notre outil d'annotation un moteur d'expansion de requêtes basé sur des règles morpho-graphématiques, VariaLog. L'une des caractéristiques essentielles de ce projet est de se dérouler dans un environnement disposant de très peu de compétences informatiques, les stratégies les plus répandues en TAL se trouvant de ce fait exclues. Nous utilisons donc AnaLog, un outil d'annotation manuelle développé précisément pour répondre à cette situation.

Abstract.

Building lexical resources for Parlanjhe, a language of the Oïl area.

The topic of this paper is the constitution of lexical resources for poitevin-saintongeais, a regional language of western France, one of the “under-resourced languages”. Since 2006, the TELPOS Project (Electronic Texts in poitevin-saintongeais) has helped constitute a database of texts, mainly literary texts, characterized by a wide variation, both diatopic and diachronic (the first text dates back to the XVIth century). The “parlanjhe” is a language of the oïl area, and therefore morphologically close to French. We chose to adapt the lexical resources at our disposal for French by adding to our annotation tool an engine called VariaLog, to expand queries on the basis of morpho-graphemic rules. One of the crucial characteristics of this project is that it is evolving in an environment with very little computer knowhow. The most widespread strategies prevailing in computer linguistics are therefore out of reach. This is why we use AnaLog, a manual annotation tool developed precisely to answer the needs of such situations.

Mots-clés : AnaLog, VariaLog, linguistique de corpus, annotation manuelle, annotation morpho-sytaxique, création de ressources lexicales, poitevin-saintongeais, parlanjhe.

Keywords: AnaLog, VariaLog, corpus linguistic, manual annotation, POS tagging, building of lexical resources, poitevin-saintongeais, parlanjhe.

1 Introduction

Le présent article porte sur la constitution de ressources lexicales pour le Poitevin-Saintongeais, langue régionale (très) faiblement dotée. Depuis 2006, le projet TelPOS (Textes Electroniques en poitevin-saintongeais), a permis la constitution d'un base de données de textes (essentiellement littéraires) caractérisés par une forte variation, tant diatopique que diachronique (le premier texte date du 16e siècle). L'annotation du corpus a été envisagée dès les débuts du projet. Le Poitevin-Saintongeais est une langue d'oïl, donc morphologiquement proche du français : nous avons choisi d'adapter des ressources élaborées pour l'annotation du français du 16e siècle dans le cadre du projet des Bibliothèques Virtuelles de Tours (CESR, www.bvh.univ-tours.fr). L'une des caractéristiques essentielles de ce projet est de se dérouler dans un environnement disposant de très peu de compétences informatiques, les stratégies les plus répandues en TAL se trouvant de ce fait exclues. Nous utilisons donc AnaLog, un outil d'annotation manuelle développé précisément pour répondre à cette situation. La création des ressources lexicales se fait en cours d'annotation, AnaLog

proposant diverses fonctionnalités permettant d'incrémenter les ressources en cours de traitement. Pour bénéficier au mieux des ressources disponibles, nous avons intégré à l'outil un moteur d'expansion de requêtes basé sur des règles morpho-graphématiques, VariaLog : il permet de « reconnaître » une forme rencontrée dans un texte comme étant probablement à mettre en relation avec une forme décrite dans le dictionnaire de français standard.

Après avoir présenté le corpus et les caractéristiques linguistiques du poitevin-saintongeais, nous décrirons les choix faits pour l'étiquetage. Puis nous exposerons ensuite la méthodologie sur laquelle repose AnaLog et l'intégration du moteur d'expansion de requête.

2 Un corpus de poitevin-saintongeais

2.1 Le projet TELPOS

Au contraire de certaines des langues de France qui souffrent parfois d'un manque de production écrite, on trouve pour le parlanjhe une production écrite et littéraire ininterrompue au moins depuis le 16e siècle, marquée inauguralement par les recueils de textes intitulés *La gente poitevinerie* puis *Le Rolea* au 17e siècle. Les textes en parlanjhe sont disponibles et accessibles par l'intermédiaire des bibliothèques et des rééditions, mais, en dépit d'une bonne présence sur le net[1] ils ont néanmoins rarement fait l'objet d'une quelconque valorisation. Le projet TELPOS (Textes électroniques en poitevin-saintongeais), a vu le jour en 2006, il est porté par la MSHS de Poitiers (laboratoire FoReLL) et soutenu par la région Poitou-Charentes dans le cadre d'une dynamique générale autour de la préservation du patrimoine culturel et de sa promotion par le numérique. Il s'agissait de créer une bibliothèque virtuelle permettant de disposer des outils numériques d'exploration des textes, que ce soit à des fins d'études littéraires et linguistiques ou pour répondre à une demande sociale qui entre en résonance avec les politiques d'aménagement du territoire en matière linguistique. Il est par exemple prévu de produire un atlas régional cartographié de la variation diachronique et diatopique (Vendée, Charente-Maritime, Deux-Sèvres, la Vienne, Charente et nord de la Gironde).

2.2 Quelques caractéristiques du Parlanjhe

Y en avét prtant yin, sou l'enpire de Badinghét, qui fàetét jhamé la mort de sun gorét é qu'alét jhamé menjhàe de boudins (Lés boudins a Nicolét)

Il en était pourtant un, sous l'empire de Badinget, qui ne fêtait jamais la mort de son cochon et qui n'allait jamais manger de boudin (Les boudins de Nicolas).

Le parlanjhe est une langue d'oïl méridionale, aux confins sud-ouest du domaine, entre Loire et Gironde, au voisinage de la langue occitane. Le poitevin-saintongeais possède en outre un substrat d'oc que des états anciens de la langue attestent dans des textes du 11e par exemple, de même que la toponymie, en particulier la répartition des noms de lieux de suffixe *–ac*, résultat du suffixe celte latinisé *-acos>-acum* (le domaine de).

Il s'agit d'une langue morphologiquement proche du français, qui se distingue des autres idiomes d'oïl par quelques traits saillants, repérables d'une part dans les zones de « mots grammaticaux » et de flexion verbale, d'autre part du fait de variations graphiques transposant évolutions et variations phonétiques[2].

- Pour les premières, on peut notamment mentionner la présence d'un pronom personnel sujet de 1° et 4° *i* issue du latin *ego* : *i me di qu'ol ét rén* (*je me dis que c'est rien*), *i alun vére çheù* (*nous allons voir ça*), et celle du pronom neutre sujet *o* (*ol* devant voyelle et *ou* comme enclitique) et du pronom neutre complément d'objet direct *(z-)ou* : *o moulle* (*il pleut*), *vat-ou ?* (*est-ce que ça va ?*), *i vae z-ou dire* (*je vais le dire*). On peut aussi donner ici la conjugaison de *dounàe* (*donner*) au présent et au futur : *doune, dounes, doune, dounun, dounéz, dounant / douneraе, douneraes, dounerat, dounerun, douneréz, dounerant*.

- Pour les secondes, nous citerons la palatalisation des groupes issus du latin [p+l], [b+l], [c+l], [g+l] et [f+l] noté *pll- [pj], bll- [bj], cll- [kj], gll- [gj] et fll- [fj]* en graphie normalisée : *pllanjhe* (*calme*), *bllai* (*blé*), *cllan* (*taon*), *glla* (*glaçon*), *flla* (*fléau*). Enfin, la plupart des auteurs ont par le passé effectué des choix de notation,

[1] Citons en particulier, pour les documents oraux : Cerdo.fr, www.metive.org/documentation ; pour des documents du 13e au 20e siècle : pivetea.free.fr ; pour la production récente : http://parlanjhe.asteur.fr ; http://parlanjhe.free.fr ; clubdelanguesregionales.asso.univ-poitiers.fr.

[2] C'est là fait commun entre langues morphologiquement proches.

parfois singuliers, selon des critères morpho-phonétiques propres au parler local, au détriment de la mise en valeur de traits communs, ou plus communs, de la zone linguistique. Le traitement de la variation graphique est donc un enjeu majeur pour l'étiquetage des textes. Nous en donnons ici quelques exemples :

français (TLF 2015)	:	*ailleurs / aloi / beaucoup / domaines / fausseté / grand*
Gente Poitevinrie 1572	:	*aillours / alouay / beacot / demones / faussetez / grons*
Gente Poitevinrie 1646	:	*aillours / aloüay / beacop / demoynes / fausseti / grond*
Dictionnaire P-S Pivetea 1996	:	*allour / *aloe / beacop / deménes / faussetez / grand-grant*

Le corpus TELPOS est à ce jour constitué d'une centaine de textes, soit environ 400 000 mots, corpus que nous avons décidé d'annoter en morpho-syntaxe, en exploitant la proximité du parlanjhe et du français.

2.3 Les choix pour l'étiquetage

Notre objectif n'est pas ici de procéder à un étiquetage fin du corpus mais à (1) de procéder à une première « passe » d'annotation totale (de tous les mots) qui pose des bases validées à partir desquelles (2) on pourra acquérir une première version du lexique, (3) puis élaborer les paradigmes flexionnels qui permettront par la suite (4) d'affiner cet étiquetage : autrement dit nous avons opté pour une stratégie permettant de diminuer la complexité de la tâche d'annotation, tout en préparant un terrain facilitant l'enrichissement ultérieur. Nous nous inscrivons ici dans une campagne d'annotation construite en sous-étapes, en respectant les points de vigilance mentionnés par Fort (2012). Dans cette première phase, notre objectif est donc de procéder à une première catégorisation « fiable », c'est-à-dire que nous choisissons de remettre à plus tard la levée des zones d'ambiguïtés délicates, tout en les isolant clairement des zones où la levée de l'ambiguïté ne pose pas de problème : pour le poitevin-saintongeais (y compris en diachronie), ces zones sensibles sont assez comparables à celles que l'on rencontre pour le français (en diachronie, voire à l'époque contemporaine). De ce fait, nous avons opté pour un système distinguant différentes catégories majeures et une seule catégorie mineure pour tout ce qui concerne les « mots grammaticaux ». Nous avons retenu quatre catégories majeures « classiques » : Nom Propre – Nom – Adjectif – Verbe – Adverbe (en -ment, les autres étant traités dans la catégorie mineure, ce qui permettra de revoir dans un même temps les zones délicates de la distinction entre préposition et adverbe), auxquelles s'ajoutent deux catégories « moins classiques » qui permettent d'anticiper les problèmes réguliers de désambiguïsation. Afin de garantir une qualité constante de l'annotation et de permettre un traitement ultérieur fin de la distinction problématique entre adjectif et participe, nous avons ainsi opté pour une catégorie « adjectif-participe-gérondif », à l'instar des choix faits dans le projet PRESTO[3]. Ce choix nous amène à créer une catégorie à part pour les verbes *être-avoir*, dans la mesure où nous sommes dès lors dans une situation où la distinction entre les emplois pleins et les emplois d'auxiliaire ne sont plus systématiquement identifiables.

3 L'annotation en cours

3.1 L'outil d'annotation

Pour annoter le corpus ainsi constitué, nous avons utilisé AnaLog, outil dédié à l'exploration humaniste des textes (Lay & Pincemin, 2010), qui a pour vocation de permettre l'étude des textes en rendant possible leur annotation manuelle systématique : l'outil, pensé de façon très générale, permet à des non-spécialistes d'annoter partiellement ou totalement leur corpus avec un jeu d'étiquettes (pas limité au domaine morpho-syntaxique) dont ils décident. AnaLog occupe donc une place « à part » dans le champ des outils d'annotation manuelle : outre le fait qu'il s'agit d'un outil très léger, que tout un chacun peut prendre en main en quelques heures il se distingue par quatre caractéristiques essentielles à nos yeux, fonctionnalités qui ne sont pas disponibles à notre connaissance dans les autres outils (Fort 2012) :

1. il propose, dans un seul tableau, une visualisation en trois volets : le texte brut, l'accès aux ressources d'annotation disponibles pour chacune des formes rencontrées, et l'espace « annotation validée »
2. il autorise la création d'étiquettes ad hoc, pas exclusivement morpho-syntaxiques, éventuellement temporaires, répondant aux besoins du travail en cours
3. les annotations peuvent être créées et propagées **depuis** le concordancier, permettant d'apposer des « étiquettes de travail » à la volée : les étiquettes peuvent être posées de façon individuelle ou groupée depuis les résultats de

[3] http://presto.ens-lyon.fr/wp-content/uploads/2014/05/Étiquettes_Presto-2014-10-13.pdf

concordance. Ici, le choix est fait de ne pas se doter d'un ontologie des catégories, ni de pouvoir gérer des contraintes de cohérence au niveau du système. De tels outils, indispensables dans des « grandes campagnes » d'annotation ne s'imposent pas ici et « brident » trop la démarche exploratoire.

4. il intègre une fonctionnalité d'identification de variantes graphiques.

La conception d'AnaLog comme espace de croisement entre texte et descripteurs répond à la nécessité de confronter systématiquement le texte et la ressource d'annotation afin de valider les annotations et d'incrémenter la ressource dictionnairique (cf. figure 1).

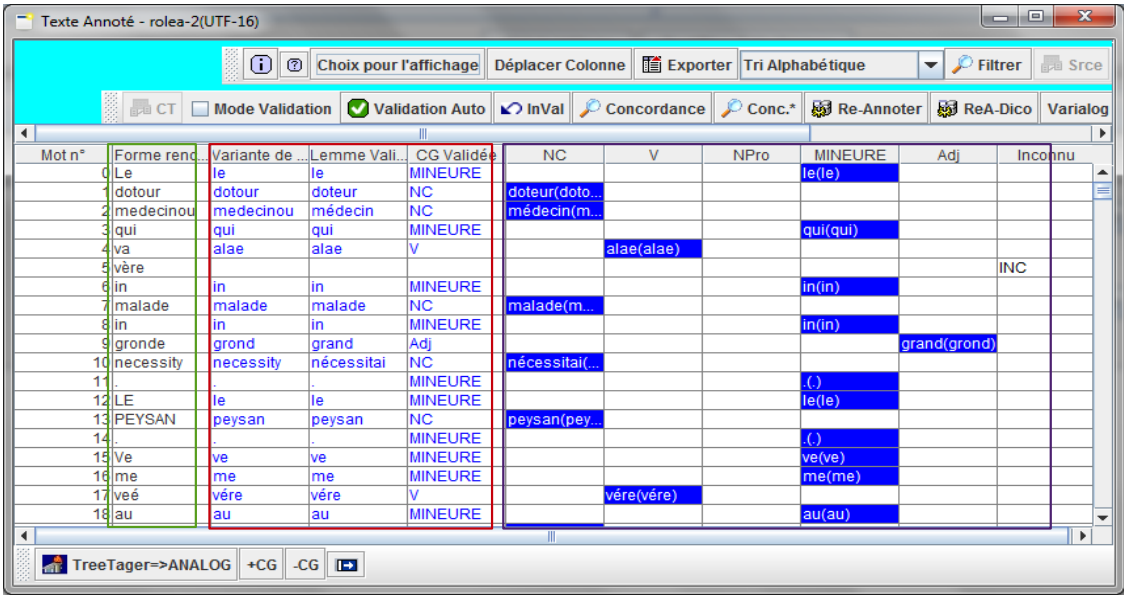


FIGURE 1: Visualisation du croisement du texte et des ressources.

Cette copie d'écran permet de visualiser les relations entre le texte dans la colonne de gauche [premier cadre vert], la ressource d'annotation dans la partie droite du tableau [troisième cadre mauve], la partie centrale [deuxième cadre rouge] donnant à voir le résultat de cette rencontre : la zone d'annotation s'affiche ici en caractères bleus ; on y trouve l'identification d'une variante de lemme, d'un lemme, et d'une catégorie grammaticale. Cette première annotation demande à être vérifiée, la ressource demande à être incrémentée : soit par un ajout de forme directement dans l'interface, pour une étiquette déjà disponible, soit en commençant par ajouter une nouvelle catégorie.

On peut partir d'un corpus brut et créer progressivement son jeu d'étiquettes comme son « dictionnaire » au cours du processus d'annotation, les formes décrites pouvant être projetées, au fur et à mesure, sur l'ensemble du corpus. On peut aussi projeter une ressource externe[4] et « voir » comment elle permet de rendre compte des phénomènes observés[5]. Son usage est donc pertinent ici, pour la création de ressources d'annotation dans un environnement « non-doté » en compétences informatiques.

[4] On peut même en projeter plusieurs, proposant une réannotation du texte par un autre dictionnaire pour compléter une ressource insuffisante. Ceci permet d'élaborer une stratégie d'annotation en plusieurs couches, fixant certaines zones avant de continuer l'annotation complémentaire. Cette façon de procéder allège le travail de désambiguïsation.

[5] On peut aussi travailler sur un corpus pré-annoté avec un autre outil pour procéder à la validation par exemple, ou pour élaborer le corpus d'apprentissage. C'est le cas dans le projet ANR-DFG Presto, ou encore dans un autre projet DFG portant sur l'annotation d'une grande collection de textes du théâtre classique (en vers et en prose, theatre-classique.fr).

3.2 La création de ressources lexicales

Nous nous basons sur le fait que le poitevin-saintongeais et le français sont des langues étymologiquement proches pour créer les ressources d'annotation / les annotations des textes. Nous avons commencé par une simple projection d'un dictionnaire de français langue générale (450 000 formes) : 35,6% des formes du texte sont reconnues.

Ne disposant pas d'un environnement TAL, nous n'avons pas fait le choix d'une « induction de lexiques à partir de corpus comparables » telle qu'elle est présentée par Scherrer et Sagot (2013), mais celui d'intégrer une fonctionnalité permettant une seconde passe d'annotation avec la ressource initiale : pour augmenter le taux de reconnaissance en partant des ressources disponibles, nous avons en effet procédé à l'intégration de VariaLog[6], outil initialement conçu pour permettre, par extension de requêtes, l'identification d'une forme en contexte hétérographique. Cet outil fonctionne par l'application successive de règles de substitution (expressions régulières), générant, à partir d'une forme, toutes les graphies envisageables étant donné les variantes possibles pour chacune des séquences de lettres composant le mot : les consonnes peuvent par exemple être simples ou redoublées, [s] peut être écrit *s,ss,c,ce,sc,sç,...* Bon nombre des formes générées ne sont bien sûr pas possibles dans la langue, mais peu importe : on cherche celle qui est éventuellement attestée dans le dictionnaire. Afin de limiter la combinatoire explosive du procédé, les règles sont contextualisées au niveau du mot, afin d'exploiter les propriétés morphologiques d'apparition des séquences de lettres. Par exemple, en français de la fin du 16e siècle, le *ai* contemporain va se trouver écrit de différentes façons selon les contextes ; (1) dans le paradigme verbal, on va trouver une combinaison d'alternances *a/o* et *i/y* (*disais/disoys*) que l'on peut contextualiser : on sait ce qui peut suivre : [*,s,t,e,ent*] ; (2) dans le paradigme nominal, on va trouver des alternances *ei,è,e...* (*paine/peine*) (3) sauf dans le cas où le *ai* du français contemporain appartient à un suffixe : *aire → arie,airie,are aise → asie,ase.* Autre exemple, *d* et *t* sont substituables (ou facultatifs) en fin de mot[7], (*enfant, renard, renart*), mais pas le reste du temps[8] ; d'autre part, la flexion du pluriel permet la chute de cette consonne (*des renars*) : [(n|r)(d|t)s$ = rds,rdz,rts,rtz,rs]. On peut ainsi établir une bonne centaine de règles (plus ou moins complexes) permettant de décrire le phénomène variationnel.

Initialement élaboré pour traiter le français de la Renaissance, nous avions fait une première adaptation du logiciel afin de traiter la variation graphique de l'anglais de la Renaissance (Lay & Duchet, 2012). Pour l'anglais, les règles de contextualisation sont de type phono-graphématiques, régies par les règles accentuelles, là où elles sont morpho-graphématiques pour le français, comme nous l'avons vu plus haut. Forte de cette expérience, nous avons décidé d'intégrer le moteur de génération de variantes au processus d'identification d'une forme dans un dictionnaire[9] utilisé pour annoter les textes. Ainsi, nous pouvons, à partir d'un dictionnaire de français contemporain, gérer la variation en diachronie. Avec la variation diatopique, nous franchissons ici une nouvelle étape : pour le traitement de la langue régionale, les règles sont bien sûr adaptées pour prendre en compte les spécificités de la relation français-poitevin. Les résultats obtenus avec un nombre minimal de règles améliorent nettement la situation : on passe à 67,4% de formes reconnues.

Ce premier procédé répond aux problématiques de variation graphique liées à la dimension diachronique du corpus comme aux traces laissées dans la graphie par les variations régionales de la prononciation.

Nous avions par ailleurs mentionné deux autres phénomènes majeurs, concernant les aspects disons « grammaticaux » du lexique. Il s'agit là de zones « fermées » ou régulières[10] que nous avons donc choisi de traiter systématiquement pour simplifier le processus d'annotation et de désambiguïsation. Les deux aspects sont traités de façon différenciée : les « mots grammaticaux » sont décrits en tant que tels et intégrés au dictionnaire (augmentation ciblée de la nomenclature), les variations dues aux flexions sont intégrées au moteur de génération de formes fléchies disponible dans AnaLog. La conséquence en est que notre dictionnaire initial enrichi des formes infinitives (« variantes » identifiées) génère toutes les formes verbales sur les modèles réguliers du parlanjhe. Il nous a été ainsi rapidement possible d'atteindre un taux de couverture de 85%.

4 Conclusion

Les outils et stratégies présentés ici bénéficient de certaines fonctionnalités rendues disponibles dans la « culture TAL », sans pour autant impliquer la présence de compétences TAL dans les projets locaux menés avec peu de moyen par des

[6] Cet outil développé par le FoReLL et le Centre d'Études Supérieures de la Renaissance de Tours (UMR 7323) a été fait l'objet d'un Google Award en 2011.

[7] On peut aisément décrire les contextes finaux dans lesquels c'est possible.

[8] Pas en milieu de mot entre deux voyelles par exemple, ce qui est heureux pour notre combinatoire !

[9] Cette adaptation pose de nombreux problèmes liés au fait que l'on « change de sens » en diachronie, et que l'on connaît parfois mal l'espace de réponse pour l'identification de la forme.

[10] Et bien décrites dans les ouvrages de référence.

équipes soucieuses de préserver ce pan de culture locale qu'est la langue régionale. Pour le parlanjhe, les choses sont lancées, et des projets d'exploitation de ces ressources sont en cours, portant notamment sur l'étude des passages en parlanjhe ou en français tourangeau chez des auteurs de la première modernité (16° et 17° siècle), comme Rabelais, Ronsard, Des Perriers, Agrippa d'Aubigné. La mise en évidence de la variation linguistique régionale devrait ouvrir des perspectives sur la « naturalité » ou le sentiment d'archaïsme suscité par les langues régionales chez ces écrivains qui semblent les avoir utilisées comme ils l'auraient fait du grec ou du latin (pour le lexique) ou du jargon des « gueux et bohémiens » pour les épisodes satiriques. Les caricatures de personnages parlant poitevin-saintongeais ne font qu'accentuer le rejet de locuteurs supposés déformer le français du roi, en cours de standardisation à cette époque. Par ailleurs, tous ne sont pas obligatoirement des disciples de Ronsard : la pratique linguistique de Rabelais, en particulier, irait dans le sens tout opposé d'un « parler naturellement » qui mine le projet d'une langue « illustre » et artificielle, finalement moquée. De tels projets d'étude sont indispensables à la mise en place des projets d'annotation qui ne trouvent sinon que trop peu d'écho pour être soutenus : souvent, la démonstration de l'utilité de l'annotation comme outil de valorisation des textes reste encore à faire.

Références

BERNHARD, D., LIGOZAT A.-L. (2013). Es esch fàscht wie ditsch, oder net? étiquetage morphosyntaxique de l'alsacien en passant par l'allemand. Actes de *TALN 2013, atelier TaLaRE*, 209-220.

CASTELLANI A. (1972). L'ancien poitevin et le problème de la langue des Serments de Strasbourg. *Les dialectes de France au Moyen-Age et aujourd'hui*, 388-427.

GAUTHIER P. (1985). La langue poitevine hier et aujourd'hui, *La Boulite poitevine-saintongeaise*, n° 8, spé. Langue poitevine-saintongeaise, UPCP.

GAUTHIER P., JAGUENEAU L. (coord.) (2002). Écrire et parler poitevin-saintongeais du XVIe siècle à nos jours. *Actes du Colloque tenu à l'Université de Poitiers les 26-27 octobre 2001*, Parlanjhe Vivant-Geste éditions.

GAUTIER M. (1993). Grammaire du poitevin-saintongeais, Mougon : Geste Éditions.

JAGUENEAU L. (1999). Le parlanjhe de Poitou-Charentes-Vendée en 30 questions. La Crèche : Geste Éditions.

LAY MH., PINCEMIN B. (2010). Pour une exploration humaniste des textes : AnaLog. Actes de *JADT 2010,* 1045-1056.

LAY MH.,DUCHETJL. (2012). VariaLog : how to locate words in Early Modern Stages of French and English. Actes de *EEBO-TCP conference 2012*, Oxford, publication en ligne sur le site http://ora.ox.ac.uk/.

PIVETEA V. (1996). Dictionnaire poitevin-saintongeais, ptv-stg / français et français / ptv-stg, Mougon : Geste Editions ; 2e éd. corrigée et augmentée, 2006, Geste Editions.

RÉZEAU P. (1984). Dictionnaire des Régionalismes de l'Ouest, entre Loire et Gironde, Les Sables-d'Olonne : Le Cercle d'Or.

SCHULZE BM, CHRIST O. (1996). The CQP User's Manual, Version 1.6.
http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/CQPUserManual/HTML/

SCHERRER Y., SAGOT B. (2013). Étiquetage morphosyntaxique de langues non dotées à partir de ressources pour une langue étymologiquement proche. Actes *de TALN 2013, atelier TaLaRE*, 195-208.

VERGEZ-COURET, M. (2013). Tagging Occitan using French and Castillan Tree Tagger', Proceedings of "Less Resourced Languages, new technologies, new challenges and opportunities, *Human Language Technologies as a Challenge for Computer Science and Linguistics.* Proceedings of the 6th *Language & Technology Conference*, (Poznań: University of Poznań), 78-82.

PRESTO : jeu d'étiquettes : http://presto.ens-lyon.fr/wp-content/uploads/2014/05/Étiquettes_Presto-2014-10-13.pdf

Quand l'oral se fait entendre à l'écrit : alignement de lexiques en l'absence de normalisation graphique

Delphine Bernhard Lucie Steiblé
LiLPa, Université de Strasbourg, 14 rue Descartes, F-67084 Strasbourg Cedex
dbernhard@unistra.fr, steiblelucie@gmail.com

Résumé. Les dialectes parlés en Alsace, que l'on regroupe communément sous l'appellation « alsacien », se caractérisent par un manque de ressources numériques, qu'il s'agisse de corpus ou de lexiques. Par ailleurs, les dialectes d'Alsace sont avant tout des langues parlées dans la vie quotidienne, et leur graphie n'est pas encore complètement codifiée : une unité lexicale peut donc avoir plusieurs graphies. Ceci est un défi majeur pour la construction de ressources lexicales, car les variantes orthographiques d'une entrée lexicale doivent être identifiées. Cet article décrit une méthode pour la construction de lexiques bilingues français-alsacien qui vise à résoudre ce problème. Elle consiste à aligner des lexiques bilingues existants, en utilisant l'algorithme phonétique *Double Metaphone* afin de détecter les variantes. En outre, les mots alsaciens sont automatiquement reliés aux entrées de BabelNet, un réseau sémantique multilingue (Navigli & Ponzetto, 2012). La méthode d'alignement des lexiques atteint de bons niveaux de précision, ce qui permet la construction automatique de ressources, avec une intervention humaine limitée à quelques corrections. La principale originalité de ce travail est qu'il ne vise pas la normalisation, qui consisterait à transformer les variantes orthographiques en une norme donnée. Par ailleurs, au lieu d'une simple liste de mots bilingues, les liens vers BabelNet fournissent une couche sémantique supplémentaire reliant les entrées à des sens lexicaux. Enfin, nous utilisons les alignements obtenus pour faire une comparaison entre observations réalisées sur la langue orale et les graphies relevées dans les lexiques.

Abstract.

From Spoken to Written : Lexicon Alignment in the Absence of an Orthographic System

The dialects spoken in Alsace, which are commonly grouped under the name "Alsatian", are characterized by a lack of digital resources, whether corpora or lexicons. Moreover, the Alsatian dialects are primarily spoken in everyday life, and their spelling is not yet completely codified : a given lexical unit can have multiple spellings. This is a major challenge for building lexical resources because alternative spellings of a lexical entry must be identified. This article describes a method for building French-Alsatian bilingual lexicons that aims to solve this problem. It consists in aligning existing bilingual lexicons, using the phonetic algorithm *Double Metaphone* to detect variants. In addition, the Alsatian words are automatically linked to entries in Babelnet, a multilingual semantic network (Navigli & Ponzetto, 2012). The lexicon alignment method achieves good levels of precision, which allows the automatic construction of resources with limited human intervention. The main originality of this work is that it does not target normalization, which would transform the spelling variants to a given standard. Moreover, instead of a simple list of bilingual words, links to Babelnet provide an additional semantic layer which connects the lexical items to senses. Finally, we use the alignments obtained to perform a comparison between phenomena observed in the spoken language and the written forms found in the lexicons.

Mots-clés : alignement de lexiques, variantes orthographiques, alsacien, BabelNet.

Keywords: lexicon alignment, spelling variants, Alsatian, BabelNet.

1 Introduction

La constitution de ressources lexicales est l'une des étapes obligatoires pour le développement d'outils du traitement automatique des langues (TAL). Cette tâche peut sembler triviale, mais dans les faits elle peut être rendue complexe par l'absence de convention orthographique, dans le cas des langues orales ou faiblement normalisées à l'écrit. Dans cet article, nous nous concentrons sur le cas précis des dialectes parlés en Alsace. Les dialectes alsaciens appartiennent aux groupes alémanique et francique (d'où l'utilisation du pluriel pour « les dialectes ») et se rapprochent de fait des dialectes parlés dans les régions limitrophes d'Allemagne et de Suisse (Huck *et al.*, 2007). Selon une étude récente, 43% de la

population alsacienne parle encore le dialecte régional (OLCA / EDinstitut, 2012). Cependant, la proportion de locuteurs alsaciens diminue régulièrement depuis les années 1960, au profit de la langue française. En outre, les dialectes alsaciens sont avant tout des langues parlées dans la vie quotidienne et leur graphie n'est donc pas encore complètement codifiée, ce qui complique toute tentative de développement de ressources et outils pour le traitement automatique des langues. Il y a eu quelques initiatives récentes visant à définir des conventions orthographiques pour l'alsacien. Le système ORTHAL (Zeidler & Crévenat-Werner, 2008) se réfère à l'orthographe allemand standard tout en permettant la transcription des phénomènes qui sont spécifiques aux dialectes alsaciens. Le système GRAPHAL-GERIPA (Hudlett & Groupe d'Etudes et de Recherches Interdisciplinaires sur le Plurilinguisme en Alsace et en Europe, 2003) définit un ensemble de règles pour aller du son au graphème. Cependant, il est difficile d'estimer la diffusion effective et l'utilisation de ces systèmes. Par ailleurs, ils accueillent la variation pour les différentes variantes géolinguistiques rencontrées en Alsace et ne garantissent donc pas une orthographe unique pour la forme d'un mot donné. Enfin, ils ne s'appliquent pas aux écrits plus anciens.

Du point de vue du traitement automatique des langues, il n'existe pas à l'heure actuelle de lexique informatisé pour l'alsacien (Leixa *et al.*, 2014), qui pourrait être utilisé pour diverses applications. Un lexique de ce type devrait idéalement comporter différents types d'informations : liste de formes attestées – pour la reconnaissance optique de caractères –, catégories grammaticales – pour l'étiquetage morphosyntaxique –, et traduction en français – pour l'aide à la lecture, les lecteurs étant souvent plus à l'aise en français qu'en alsacien –.

Pour résumer, les dialectes alsaciens posent plusieurs défis importants pour le TAL :
– Il n'existe pas de convention orthographique utilisée de manière systématique à l'écrit ;
– Le dialecte alsacien est en fait un continuum de dialectes, avec des variantes géolinguistiques tant au niveau lexical qu'au niveau de la prononciation ;
– Il n'y a pas encore de ressources lexicales numériques pour l'alsacien.

Dans cet article, nous présentons une méthode de construction de ressources lexicales numériques pour les dialectes alsaciens qui consiste à aligner plusieurs lexiques français-alsacien bilingues. Par ailleurs, nous souhaitons nous insérer dans le mouvement récent appelé *Linked Open Data* (LOD) qui vise a constituer de grandes ressources linguistiques multilingues inter-reliées. Toutes les langues ne sont pas encore couvertes, en raison du manque d'informations disponibles pour les langues faiblement dotées. Cependant, le LOD constitue une formidable opportunité pour accroître la visibilité des langues minoritaires ou régionales, si tant est qu'elles peuvent y être incorporées. Par ailleurs, le LOD permet d'accéder à de nombreuses ressources lexicales et sémantiques qui pourraient bénéficier au traitement automatique des dialectes alsaciens (définitions, liens sémantiques, etc.). Nous proposons donc d'associer des mots alsaciens à Babelnet, un réseau sémantique multilingue relié au *Linguistic Linked Open Data* (Navigli & Ponzetto, 2012).

Notre méthode s'appuie sur les observations suivantes :
– Les conventions orthographiques adoptées dans les lexiques alsaciens sont très variables, et donc la forme de citation d'une unité lexicale en alsacien peut être représentée par plusieurs graphies [1]. Ces diverses formes peuvent être considérées comme des variantes car elles sont proches phonétiquement et correspondent à la même unité lexicale. Par ailleurs, la plupart des formes de mots alsaciens sont semblables à leur traduction en allemand standard et même parfois en anglais.
– Une unité lexicale en alsacien peut avoir plusieurs traductions en français. Ces traductions peuvent être des synonymes, mais également correspondre à différents sens du mot. Cela complique l'alignement entre les deux langues, ainsi que l'alignement avec une ressource comme Babelnet, dont les unités correspondent à des concepts.

Nous abordons ces questions comme suit :
– Nous proposons d'utiliser une variante d'un algorithme phonétique, *Double Metaphone*, adapté aux dialectes alsaciens, afin d'identifier les variantes orthographiques. L'algorithme prend également en compte l'orthographe de l'allemand standard et de l'anglais afin de trouver des mots apparentés dans ces diverses langues.
– Nous utilisons des ressources externes pour obtenir des informations sur les synonymes dans la langue française et des traductions en allemand et en anglais.

L'article est organisé comme suit : la section suivante récapitule les travaux antérieurs sur l'identification des variantes orthographiques et l'alignement des ressources lexicales. La section 3 détaille les ressources lexicales utilisées dans notre travail. Les méthodes d'alignement sont présentées dans la section 4, qui comprend également une évaluation des alignements obtenus sur la base d'un dictionnaire multilingue publié. Enfin, nous faisons une comparaison entre observations faites sur la langue orale et les graphies relevées à l'écrit, sur la base des alignements obtenus de manière automatique.

1. Dans cet article, la notion d'unité lexicale renvoie à un lexème, dans le sens de Bauer (2003) : « Un lexème est un mot du dictionnaire, une unité abstraite du vocabulaire. Il est réalisé (...) par des mots-formes (*word forms*), de telle sorte que le mot-forme représente le lexème et toutes les flexions (...) qui sont nécessaires. (...) La forme de citation d'un lexème est le mot-forme appartenant au lexème qui est conventionnellement choisi pour nommer le lexème dans les dictionnaires et autres. » (notre traduction)

2 État de l'art

2.1 Identification de variantes orthographiques

Le problème des graphies non standard se rencontre pour différents types de textes, comme par exemple les données issues du Web (en particulier le Web 2.0), les textes correspondant à des états anciens de la langue, et les textes écrits dans des langues qui sont principalement orales et qui n'ont pas de système orthographique. La grande majorité des méthodes consiste à normaliser vers une langue cible, c'est-à-dire, transformer une variante minoritaire en une norme donnée. Par exemple, Scherrer (2008) utilise la distance orthographique de Levenshtein et des transducteurs stochastiques afin de transformer les formes dialectales du Suisse allemand en allemand standard. Hulden *et al.* (2011) présentent deux méthodes qui apprennent automatiquement les transformations d'une forme dialectale vers la forme standard en utilisant un corpus parallèle pour la langue basque et le dialecte basque labourdin. La première méthode s'appuie sur un outil existant, lexdiff (Almeida *et al.*, 2010), qui détecte les différences orthographiques. Ces différences sont transformées en règles de remplacement et compilées sous forme de transducteurs. La deuxième méthode est inspirée par la PLI (programmation logique inductive) et essaie de sélectionner le meilleur ensemble de règles de remplacement, en utilisant des exemples à la fois positifs et négatifs. Dans le contexte de la traduction automatique statistique pour la paire de langues arabe-anglais, Salloum & Habash (2011) décrivent une méthode à base de règles pour générer des paraphrases de l'arabe dialectal en arabe standard. Pour les variantes linguistiques historiques, Porta *et al.* (2013) proposent une méthode pour mettre en correspondance les formes historiques avec leurs homologues modernes. L'approche est basée sur un transducteur de Levenshtein et un transducteur linguistique encodant des règles de réécriture des sons.

Dans l'ensemble, les méthodes de normalisation considèrent que les dialectes ou les formes de mots historiques sont non-standard et doivent être transformées dans des formes contemporaines d'une langue bien dotée en ressources. Même si cette hypothèse est logique dans de nombreux cas, notamment pour faciliter le traitement ultérieur par les outils de TAL, ce n'est pas la seule solution. Par exemple, Dasigi & Diab (2011) présentent un algorithme de clustering qui vise à regrouper les variantes orthographiques dialectales qui correspondent au même mot. Ce type d'approche est particulièrement pertinent dans notre contexte, car il ne normalise pas nécessairement les variantes dialectales. En effet, la normalisation n'est pas souhaitable dans le cas des dialectes alsaciens pour plusieurs raisons. Tout d'abord, il n'y a pas consensus sur la norme de scripturalisation des dialectes alsaciens et il est donc difficile de décider quelle forme doit prévaloir. En outre, même si les dialectes alsaciens sont étroitement liés à l'allemand, qui pourrait être considéré comme le standard, il existe un certain nombre de différences lexicales (notamment des emprunts au français (Matzen, 1985)) et syntaxiques (voir par exemple (Kleiber & Riegel, 1998)) qui doivent être prises en compte. Ajouté à cela, considérer l'allemand comme la norme pour les dialectes alsaciens est une question très sensible du point de vue sociolinguistique, voire politique [2]. Compte tenu de toutes ces raisons, notre méthode ne cherche pas à normaliser les variantes graphiques mais conserve leur diversité en considérant des groupes (ou *clusters*) de variantes comme des entrées du lexique.

2.2 Alignement de ressources lexicales

L'objectif principal de notre travail n'est pas seulement d'identifier les variantes orthographiques, mais aussi d'aligner les entrées issues de différents lexiques bilingues et de mettre en correspondance ces alignements avec les concepts d'un réseau sémantique. Beaucoup de travaux ont été consacrés récemment à l'alignement de ressources collaboratives, comme Wikipedia, et de bases de connaissances lexicales plus classiques, comme WordNet. Niemann & Gurevych (2011) détaillent une méthode pour l'alignement des sens des entrées de WordNet et Wikipedia, qui a ensuite été utilisée pour la ressource lexicale sémantique UBY (Gurevych *et al.*, 2012). La méthode repose sur l'apprentissage automatique afin de classer les alignements comme valides ou non valides. La similitude des sens candidats est calculée sur la base d'une représentation "sac de mots" des sens, puis fournie au classifieur. Pour la ressource UBY, des alignements translingues sont induits de la même manière, en traduisant tout d'abord automatiquement les représentations textuelles des sens. Navigli & Ponzetto (2012) proposent une méthode pour relier les pages Wikipédia aux synsets de WordNet, qui a été utilisée pour la construction de la ressource BabelNet. La méthode applique plusieurs stratégies en séquence. En particulier, elle réutilise une technique proposée dans le cadre de la désambiguïsation lexicale qui consiste à définir un contexte de désambiguïsation pour chaque page Wikipedia et chaque sens dans WordNet. Le contexte utilisé est un ensemble de mots obtenus à partir des informations fournies dans les ressources (par exemple, les noms des pages, les liens, les redirections et les catégories dans Wikipedia ; les synonymes, hyperonymes / hyponymes, gloses dans WordNet). Un score de similarité

2. Il est vrai que cet argument n'a pas beaucoup de poids du point de vue pratique pour les outils de TAL, mais il s'ajoute aux autres.

peut alors être calculé sur la base de ce contexte.

Quand il n'y a aucune ressource lexicale dans une langue donnée, la traduction automatique des ressources d'une autre langue est souvent la meilleure option, en terme de coût de construction. Dans ce cas, une ressource existante est étendue avec les lexicalisations d'une autre langue et la structure est conservée. Le WOLF (Wordnet Libre du Français) a été construit par Sagot & Fišer (2008) en utilisant le Princeton WordNet et plusieurs ressources multilingues. Les principales hypothèses qui sous-tendent leur approche sont que les différents sens d'un mot ambigu dans une langue correspondent souvent à différentes traductions dans une autre langue, et les mots qui sont traduits par le même mot dans une autre langue ont souvent des significations similaires. Ils appliquent ces idées en recueillant un lexique multilingue comportant 5 langues à partir d'un corpus parallèle et en assignant le synset le plus probable à chaque entrée du dictionnaire, en s'appuyant sur les intersections entre les synsets associés à chaque mot non-français du lexique dans le Princeton WordNet ou dans les wordnets du projet BalkaNet. Hanoka & Sagot (2012) ont étendu la ressource WOLF en utilisant une nouvelle approche qui s'appuie sur un grand grand graphe de synonymes et de traductions construit à partir de Wikipedia et de Wiktionary. Le graphe est interrogé avec les littéraux de wordnets multilingues alignés pour obtenir le meilleur candidat de traduction, en utilisant à la fois la traduction et des relations de traduction inverse.

Dans notre travail, nous appliquons également l'idée d'étendre une ressource lexicale sémantique existante (BabelNet) avec des lexicalisations d'une autre langue, à savoir l'alsacien. Nous utilisons le français comme langue pivot pour obtenir une mise en correspondance entre les variantes alsaciennes et Babelnet. En outre, nous exploitons la proximité entre l'alsacien, l'allemand et l'anglais afin d'enrichir les vecteurs de caractéristiques et effectuer la désambiguïsation.

3 Ressources

3.1 Lexiques bilingues français-alsacien

Nous avons récupéré trois lexiques bilingues français-alsacien disponibles sur le Web :
– OLCA : les lexiques produits par l'OLCA (Office pour la Langue et la Culture d'Alsace) [3]. Ces lexiques sont spécifiques à des domaines particuliers (l'artisanat, l'automobile, la bière, les courses, l'équitation, le football, les livres, la médecine, la météo, la nature, la petite enfance, la pêche, la pharmacie, le vélo, la vigne) et fournissent dans certains cas des variantes pour les départements alsaciens du Bas-Rhin et du Haut-Rhin ;
– WKT : un lexique extrait d'une page utilisateur du Wiktionnaire [4] ;
– ACPA : un lexique bilingue disponible sur la page Web d'une association locale [5].
Les lexiques contiennent essentiellement des lemmes, ainsi que quelques expressions. Par ailleurs, ces lexiques, bien que numériques, ne sont pas disponibles dans un format standard. Ils ont été pré-traités avec des analyseurs spécifiques pour extraire les paires de mots français-alsacien. Lorsqu'elles sont disponibles, les informations sur la partie du discours sont conservées [6]. Sinon, nous avons utilisé deux heuristiques pour trouver la partie du discours : (a) utilisation du TreeTagger français (Schmid, 1994) pour obtenir la catégorie des mots simples français [7] ; (b) pour les noms, vérification de la présence d'un déterminant à côté de la forme alsacienne.

La Table 1a répertorie le nombre d'entrées dans la partie française des lexiques après pré-traitement. Le tableau montre que la couverture des différentes parties du discours est inégale, et que les lexiques se concentrent principalement sur les noms, les verbes et les adjectifs.

Les lexiques suivent différentes conventions orthographiques comme le montre la Table 1b [8], qui énumère les traductions trouvées dans les lexiques pour plusieurs mots. Beaucoup de traductions dans la table sont en fait des variantes graphiques du même mot alsacien (par exemple "Kràb" et "Kràpp"). Toutefois, ces variantes graphiques peuvent être très dissemblables, si on ne considère que les caractères utilisés.

3. http://www.olcalsace.org/
4. http://fr.wiktionary.org/wiki/Utilisateur:Laurent_Bouvier/alsacien-fran\%C3\%A7ais
5. Compilé par André Nisslé, http://culture.alsace.pagesperso-orange.fr/dictionnaire_alsacien.htm
6. Nous avons utilisé la liste de catégories suivante : verbe, adjectif, adverbe, préposition, locution, conjonction, pronom, interjection, nom propre, participe passé, déterminant, abréviation, nom (féminin, masculin, neutre, pluriel).
7. Nous utilisons le module TreeTaggerWrapper de Laurent Pointal disponible à http://perso.limsi.fr/pointal/dev:treetaggerwrapper. Nous préférons l'utilisation de l'étiqueteur à un lexique morphosyntaxique du français car il donne l'étiquette la plus probable, alors que dans un lexique morphosyntaxique toutes les étiquettes sont équiprobables.
8. Voir également la Table 4, p. 9.

	OLCA	WKT	ACPA
adjectif	224	122	1 898
adverbe	14	49	295
déterminant	1	20	15
nom	5 106	1 049	15 770
participe passé	63	59	476
pronom	1	38	47
verbe	445	292	3 017
catégorie indéterminée	943	393	2 015
TOTAL	6 797	2 022	23 533

(a) Nombre de mots français dans les lexiques français-alsacien.

Français	**corbeau**	**jambe(s)**	**grenier**
Anglais	crow	leg	attic
Allemand	Rabe	Bein	Dachboden
ACPA	Kräje Kràbb	Bai Unterschankel	Behna **Behn** Ästrich Dàchbooda
WKT	Grâb Kràpp **Ràmm**	**Bein** **Baan**	**Behn** Behni Bhena Kàscht Späicher Spicher
OLCA	Kràb **Ràmm**	**Bein** Bei **Baan**	Hejbodde Dàchstüel **Behn**

(b) Exemples de traductions trouvées dans les lexiques. Les variantes qui se retrouvent de manière identique dans au moins deux lexiques sont en gras.

En plus des lexiques bilingues, nous avons également utilisé deux réseaux sémantiques : JeuxDeMots et Babelnet.

3.2 JeuxDeMots

JeuxDeMots (Lafourcade, 2007) est un réseau lexical français disponible gratuitement et construit à l'aide de jeux en ligne [9]. Nous avons utilisé la version datée du 12 Juin 2014 [10], qui contient 178 569 occurrences de la relation de synonymie (le réseau contient également de nombreux autres types de relations, par exemple association, domaine, hyperonymie, hyponymie, etc.). JeuxDeMots est utilisé pour relier des entrées dont les traductions en français sont synonymiques, comme par exemple 'drăckĭg' - *sale* et 'trackig' - *malpropre*.

3.3 BabelNet

BabelNet (Navigli & Ponzetto, 2012) est un réseau sémantique multilingue, qui intègre des données issues de WordNet et Wikipedia, entre autres. Babelnet est composé de synsets, qui correspondent à des concepts avec des lexicalisations en plusieurs langues. Les lexicalisations multilingues ont été obtenues soit grâce aux liens inter-langues de Wikipédia ou à la traduction automatique. Nous avons utilisé la version 2.5 de Babelnet [11].

4 Alignement des lexiques

Dans cette section, nous présentons notre méthode pour aligner les lexiques. Elle repose sur une adaptation de l'algorithme phonétique *Double Metaphone* aux dialectes alsaciens.

4.1 *Double Metaphone* pour les dialectes alsaciens

Compte tenu de l'absence de convention orthographique, ainsi que des différences dues à la variation géolinguistique, il n'est pas possible d'aligner les entrées issues de différents lexiques en fonction de la similarité graphique des formes (Considérons par exemple "Grâb" et "Kràbb" de la table 1b, qui n'ont que deux caractères communs : 'r' et 'b'). Afin de résoudre ce problème, nous avons développé un algorithme *Double Metaphone* pour les dialectes alsaciens. *Double Metaphone* (Phillips, 2000) a été proposé à l'origine pour la recherche d'information, afin de trouver des noms orthographiés

9. Voir `http://www.jeuxdemots.org`
10. Disponible sur `http://www.lirmm.fr/~lafourcade/JDM-LEXICALNET-FR`
11. Disponible à `http://www.babelnet.org`

différemment, mais faisant référence à la même entité. *Double Metaphone* appartient à la classe des algorithmes phonétiques, car il transforme la chaîne d'entrée en une clé qui est identique pour les mots qui sont prononcés d'une manière similaire. Par exemple, la clé metaphone est STFN pour les trois noms suivants : "Stephan", "Steven" et "Stefan". Afin de prendre en compte les ambiguïtés, *Double Metaphone* retourne en fait deux clés dans certains cas. *Double Metaphone* a par exemple été utilisé pour la normalisation de textes du Web 2.0 (Mosquera *et al.*, 2012).

Les transformations *Double Metaphone* implémentées pour l'alsacien sont basées sur des transformations proposées à l'origine pour l'anglais [12] et une analyse de nos lexiques. Nous avons constitué un jeu de test comprenant 141 formes alsaciennes et les clés métaphones attendues, avec diverses variantes de la même unité lexicale, afin de vérifier que les clés métaphones sont bien identiques dans ce cas. Nous avons également pris en compte l'allemand standard, afin d'obtenir des clés identiques pour les mots apparentés (cognats) allemands et alsaciens. La table 2 donne quelques exemples des clés metaphone obtenues pour plusieurs mots alsaciens et allemands.

Mot	Traduction en français	Clé metaphone 1	Clé metaphone 2
Schloofwàga	wagon-lit	XLFVK	XLFVY
Schlofwaawe		XLFVV	XLFVY
Rüejdàà	jour de repos	RT	/
Rüaijtààg		RTK	RT
beschtadiga	confirmer	PXTTK	PXTTY
Uffschtànd	insurrection	AFXTNT	/
Iwereinsschtimmung	concordance	AFRNXTMNK	AVRNXTMNK
bestätigen	confirmer	PXTTK	/
Aufstand	insurrection	AFXTNT	/
Übereinstimmung	concordance	APRNXTMNK	AVRNXTMNK

TABLE 2: Exemples de clés metaphone. Les mots alsaciens sont dans la partie supérieure de la table, et les mots allemands sont dans la partie inférieure.

4.2 Méthode d'alignement

Notre premier objectif est d'aligner les entrées dans plusieurs lexiques bilingues français-alsacien. Dans une première étape, toutes les entrées des trois lexiques utilisés sont ajoutées à un graphe. Les nœuds correspondent aux mots alsaciens et à leurs traductions en français. Les mots alsaciens sont connectés à leurs traductions en français dans les lexiques par une arête. En outre, deux mots alsaciens sont reliés par une arête si toutes les conditions suivantes sont remplies :

1. ils ont la même traduction en français ;
2. ils ont une clé métaphone en commun ;
3. ils appartiennent à la même partie du discours [13].

Nous utilisons également des informations obtenus à partir des ressources décrites à la section 3 afin d'assouplir la condition 1. La liste des synonymes français issue de JeuxDeMots est utilisée pour connecter deux mots alsaciens qui ont des traductions françaises synonymes dans cette ressource. Les sens français de BabelNet sont utilisés de la même manière que les synonymes de JeuxDeMots, pour connecter les mots alsaciens qui ont des traductions françaises ayant le même sens.

4.2.1 Alignement des variantes alsaciennes

Après la construction du graphe, les variantes alsaciennes sont regroupées. Les formes qui sont des variantes sont récupérées par la détection de composantes connexes dans le sous-graphe contenant uniquement les mots alsaciens. La figure 1 montre une portion du graphe initial. Les traductions en français, allemand et anglais sont également présentées. Dans le sous-graphe formé par les mots alsaciens, il y a trois composantes connexes : (1) ["Winkäller", "Winkeller", "Winkaller"], (2) ["Wikaller"] et (3) ["Kaller"]. Les formes "Winkäller", "Winkeller" et "Winkaller" sont donc regroupées dans un cluster et considérées comme des variantes graphiques de la même unité lexicale.

12. Notre implémentation de Double Metaphone repose sur un module Python existant : `http://www.atomodo.com/code/`

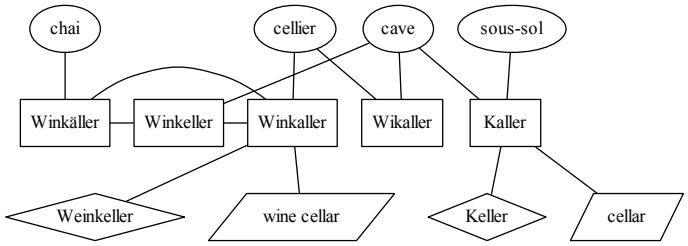


FIGURE 1: Vue simplifiée d'un sous-graphe. Les mots français figurent dans des ellipses, les mots alsaciens dans des rectangles, les mots anglais dans des parallélogrammes et les mots allemands dans des losanges.

4.2.2 Mise en correspondance avec les synsets de BabelNet

Notre deuxième objectif est de mettre en correspondance les variantes alsaciennes alignées avec les synsets de Babelnet. Par exemple, en prenant l'exemple de la figure 1, le cluster formé par ["Winkäller", "Winkeller", "Winkaller"] doit être mis en correspondance avec le synset ayant pour identifiant `bn:00017041n` (voir Figure 2).

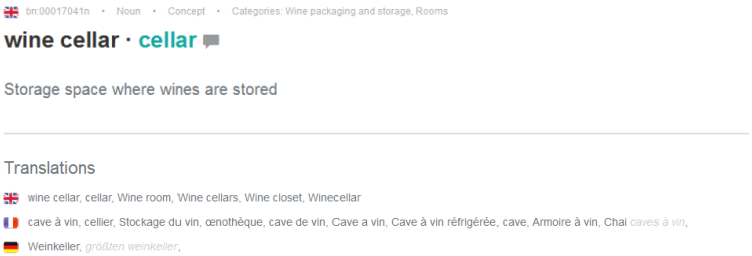


FIGURE 2: Synset `bn:00017041n` dans l'interface de recherche de BabelNet

La mise en correspondance se fait en calculant la similarité cosinus entre des représentations "sac de mots" binaires des synsets de BabelNet et des variantes alsaciennes alignées. Dans le cas le plus simple, la représentation utilisée pour les synsets de BabelNet se compose de leurs lexicalisations français. Les variantes alsaciennes sont représentées par leurs traductions en français : dans l'exemple de la Figure 1, le cluster formé par ["Winkäller", "Winkeller", "Winkaller"] sera représenté par le sac de mots ["chai", "cellier", "cave"]. Les représentations sac de mots peuvent être étendues en utilisant les traductions disponibles dans BabelNet. En effet, il a été montré que l'utilisation des traits multilingues a un effet positif sur la tâche de désambiguïsation (Banea & Mihalcea, 2011). Cependant, il faut éviter l'ambiguïté lorsque l'on sélectionne les traductions en anglais et en allemand pour les mots alsaciens. Cette question a été abordée dans les travaux sur l'acquisition de dictionnaires bilingues pour une paire de langues en utilisant une troisième langue comme un pivot : dans notre cas, le français est la langue pivot, l'alsacien la langue source et l'allemand et l'anglais les langues cibles. Plusieurs méthodes ont été proposées, qui reposent principalement soit sur la structure des lexiques bilingues disponibles soit sur la similarité distributionnelle (Salloum & Habash, 2011; Tanaka & Umemura, 1994). Dans notre cas particulier, nous exploitons la proximité entre l'alsacien et l'allemand, et, à un degré moindre, l'anglais. A partir des traductions en français, les traductions en allemand et / ou en anglais sont ajoutées aux représentations sac de mots des clusters de variantes alsaciennes si les traductions et les mots alsaciens partagent une de leurs clés metaphone. Cette contrainte effectue une sorte de désambiguïsation et assure que seules traductions valides sont sélectionnées. Ainsi, dans l'exemple de la figure 1, le mot allemand "Weinkeller" et le mot anglais "wine cellar" seront ajoutés aux sacs de mots.

`double-metaphone/metaphone.py/view`

13. Les participes passés et les adjectifs sont considérés comme faisant partie de la même catégorie.

4.3 Évaluation et résultats

Afin d'évaluer notre méthode, nous avons produit manuellement 107 alignements de référence entre les lexiques et BabelNet. Dans ce but, nous avons choisi au hasard des entrées d'un dictionnaire multilingue français-allemand-anglais-alsacien (Adolf, 2006). Ce dictionnaire présente plusieurs avantages pour l'évaluation : plusieurs variantes orthographiques sont généralement proposées pour chaque entrée alsacienne ; les traductions en français, allemand et anglais sont fournies, facilitant ainsi la mise en correspondance avec BabelNet ; enfin, le dictionnaire se concentre sur des mots alsaciens qui sont très semblables aux mots allemands et anglais correspondants. S'il est vrai que cela induit un biais dans l'évaluation pour les configurations où les lexiques anglais et allemands sont utilisés, cela permet d'avoir une idée des performances maximales qu'il est possible d'atteindre, et qui seront vraisemblablement inférieures sur l'ensemble du lexique. Pour les données d'évaluation, nous avons vérifié que les entrées se retrouvent dans au moins un des trois lexiques bilingues utilisés (OLCA, WKT et ACPA). En outre, nous avons sélectionné le meilleur synset correspondant de Babelnet. Quand il n'était pas possible de décider, au plus trois synsets de Babelnet ont été choisis.

L'alignement des variantes est évalué en terme de précision, rappel et F-mesure. Pour chaque cluster de mots alsaciens tels qu'une des traductions en français se trouve dans les données d'évaluation, nous mesurons l'intersection entre les alignements automatiques du cluster et les variantes alsaciennes dans les données de référence dans comme des vrais positifs (VP). Les variantes automatiquement alignées qui ne sont pas dans les données de référence sont considérées comme des faux positifs (FP), tandis que celles de la référence qui ne sont pas dans les alignements produits sont considérées comme des faux négatifs (FN). Par exemple, pour le cluster ['Schekbeeschel'/ACPA, 'Scheckbiechel'/Adolf(2006)] [14] (*carnet de chèque* en français), l'alignement de référence est ['Schäckbiachla'/ACPA, 'Schekbeeschel'/ACPA, 'Scheckbiechel'/Adolf(2006)]. Dans ce cas, VP=2, FP=0 et FN=1. Ensuite, la précision (P), le rappel (R) et F-mesure (F) sont calculés comme suit :

$$P = \frac{VP}{VP + FP} \quad ; \quad R = \frac{VP}{VP + FN} \quad ; \quad F = \frac{2 \cdot P \cdot R}{P + R}$$

La mise en correspondance avec BabelNet est également évaluée en termes de précision, rappel et F-mesure. Les synsets de BabelNet étant ordonnés selon la similarité cosinus, nous prenons en compte tous les synsets qui ont le même cosinus au rang 1. Les résultats de l'évaluation pour les différents paramètres sont détaillés dans la table 3. La *baseline* correspond à l'absence de ressources externes. + JDM indique que les synonymes de JeuxDeMots ont été utilisés. + BN indique que BabelNet a été utilisé, avec les lexicalisations en français (FR), allemand (DE) ou en anglais (EN).

	Alignement des lexiques			**Alignement avec BabelNet**		
	P	**R**	**F**	**P**	**R**	**F**
baseline	0,99	0,69	**0,81**	0,18	0,44	0,26
+ BN FR	0,94	0,70	0,80	0,21	0,47	0,29
+ JDM	0,97	0,70	**0,81**	0,18	0,44	0,26
+ BN FR & DE	0,94	0,70	0,80	0,41	0,49	0,45
+ BN FR & EN	0,94	0,70	0,80	0,27	0,50	0,35
+ BN FR, DE & EN	0,94	0,70	0,80	**0,45**	**0,51**	**0,48**
+ JDM + BN FR & DE	0,92	0,70	0,80	0,39	0,48	0,43
+ JDM + BN FR, DE & EN	0,92	0,70	0,80	0,44	0,50	0,47

TABLE 3: Résultats de l'évaluation

Dans l'ensemble, les résultats pour les alignements des variantes issues de différents lexiques sont stables : l'utilisation des ressources externes conduit à une légère baisse de la précision qui est compensée par une très légère hausse du rappel. En outre, le rappel est toujours inférieur à la précision. Pour la mise en correspondance avec les synsets de Babelnet, l'utilisation de traductions en allemand et, à un degré moindre, en anglais, conduit à des améliorations. Dans ce cas, moins de faux positifs sont détectés, parce que les mots allemands et anglais fournissent un contexte de désambiguïsation qui aide à identifier le synset correct. Les résultats peuvent sembler décevants, avec une F-mesure culminant à 0,48. Cependant, le mode d'évaluation est assez strict, car il permet l'alignement avec un seul synset de BabelNet dans la plupart des cas. Les synonymes fournis par JDM ont un effet légèrement négatif sur la performance, très certainement parce que les ensembles de synonymes dans cette ressource sont différents de ceux que l'on trouve dans Babelnet. Le rappel inférieur obtenu

14. Nous indiquons également le lexique d'où est issu la variante après le caractère '/'.

pour les alignements de variantes dans les lexiques est principalement dû à la contrainte qui exige des clés metaphone identiques. Dans certains cas, des variantes ont différentes clés (par exemple “Chilche” - KLX / XLX et “Kirche” - KRX). Cela soulève également une question plus fondamentale : ces formes peuvent-elles encore être considérées comme des variantes, ou correspondent-elles à des unités différentes ? Dans notre construction des alignements de référence, nous avons regroupé les variantes que l’on trouve dans le dictionnaire multilingue, même si elles pouvaient être différentes dans certains cas. Par ailleurs, en plus des clés metaphone, d’autres mesures de similarité orthographique pourraient être utilisées pour aligner les variantes, comme cela se fait pour l’identification de cognats (Inkpen *et al.*, 2005). Ces mesures pourraient aider à l’amélioration du rappel. Certaines erreurs sont également dues à des problèmes dans la récupération des parties du discours pour les entrées de dictionnaire ambiguës. Comme l’une des conditions d’alignement requiert des parties du discours identiques, ces entrées ne sont pas considérées comme des variantes.

Comme le montrent les résultats, l’ajout d’informations multilingues permet d’améliorer la mise en correspondance avec les synsets de BabelNet. Pour le moment, les traductions en allemand et en anglais sont choisies en fonction de leurs clés metaphone, ce qui conduit à des traductions manquantes pour certains “sacs de mots”. À l’avenir, ceci pourrait être amélioré en utilisant des lexiques bilingues supplémentaires, afin d’ajouter des traductions qui ne sont pas nécessairement apparentées aux variantes alsaciennes.

5 Comparaison oral-écrit

L’alignement des variantes alsaciennes trouvées dans différents lexiques permet de réaliser une étude comparative des scripturalisations utilisées, en mettant en évidence les différences les plus fréquentes en termes de remplacements de caractères. Ces différences témoignent des difficultés rencontrées lors de la transcription du dialecte à l’écrit, tant par les lexicologues que par les locuteurs non-spécialistes. La comparaison des formes trouvées dans les lexiques permet de mettre au jour les problématiques majeures de cette mise à l’écrit. Nous avons analysé 581 différences entre les mots des lexiques, plus ou moins fréquentes (la fréquence de remplacement correspond à la colonne “rang” de la Table 4). Ces différences portent essentiellement sur le remplacement d’un graphème par un autre [15].

Rang	Remplacement	Nombre d’occurrences	Exemples
1	a → e	3 307	Wàldbeerla – Wàldbeerle
2	e → i	962	blend – blind
3	e → ä	493	Hardepfel - Hardäpfel
4	a → ä	487	Base – Bäse
5	u → ù	476	Leischtung - Leischtùng
6	a → à	358	hopla – hoplà
7	e → ì	213	Leppel - Lìppel
8	i → ì	211	Kopfkisse – Kopfkìsse
9	d → t	166	dànze – tànze
…	…	…	…
13	g → j	121	Flegel – Flejel
…	…	…	…
21	b → p	61	bolìere – polìere
…	…	…	…
23	g → k	51	stàrig - stàrik
…	…	…	…
124	f → v	8	narfig - narvig
…	…	…	…
458	s → z	2	baidersits - beiderssitz

TABLE 4: Remplacements de graphèmes dans les lexiques

15. On trouve également des remplacements de plusieurs graphèmes mais ils sont plus rares.

5.1 Les voyelles

Les mots *Wàldbeerla* et *Wàldbeerle* (fraise des bois) sont un bon exemple de variabilité : seule la voyelle finale est soumise au changement. De ce fait, ils sont donc comptabilisés dans l'entrée de remplacement « a → e ». Les variantes sur ces deux voyelles sont extrêmement nombreuses : il s'agit de la première entrée du tableau de comparaison, comprenant 3 307 items transformés. Ce chiffre conséquent est corrélé à une réalité sociophonétique : l'une des nuances vocaliques les plus connues concernant l'alsacien est une différence observable entre le Nord et le Sud de l'Alsace. L'aperture des voyelles postérieures augmente en fonction de la provenance géographique des locuteurs : selon notre exemple, *Wàldbeerla* au Sud et *Wàldbeerle* au Nord. Ainsi, les remplacements de graphèmes les plus fréquents sont en accord avec, et même soulignent un fait phonétique. Les voyelles sont en fait très soumises aux variations graphiques : les huit premières entrées des remplacements portent sur ce type de phonèmes, avec au total, 6 507 formes transformées. Ces modifications concernent majoritairement des choix purement graphiques : 1 321 modifications entre des graphèmes tels que « a » et « à », par exemple entre les formes *hoplà* et *hopla*. L'instabilité des voyelles graphiques est parfois correspondante à la variabilité constatée d'un point de vue phonétique. L'utilisation de l'algorithme *Double Metaphone* est donc particulièrement pertinente, puisque les clés utilisées pour l'alignement n'utilisent que les consonnes (voir Section 4.1).

5.2 Les consonnes

Les occlusives de l'alsacien, graphiées `p,t,k` et `b,d,g`, ne sont pas les mêmes que les phonèmes graphiés en français avec les même symboles. En effet, en français, ces consonnes s'opposent selon le trait phonologique de voisement : pour produire les sourdes `/p,t,k/`, les plis vocaux cessent de vibrer pendant l'occlusion, tandis que pour les sonores, la vibration est maintenue. En allemand, cette différence est également présente, mais tend à être remplacée en position initiale de mot et parfois en finale par une opposition d'aspiration, ou de tension (il est souvent fait référence à ce phénomène en tant que *dévoisement* ou *assourdissement*). La vibration des plis vocaux est alors absente pour les deux séries. En alsacien, le voisement n'est pas forcément pertinent pour distinguer ces consonnes, problématique connue en phonétique (Bothorel-Witz & Pétursson, 1972; Erhart, 2012; Pipe, 2014; Woehrling & Boula de Mareüil, 2005). L'analyse acoustique en lecture événementielle des signaux de parole (Steiblé, 2014) a pu apporter des lumières sur ces consonnes, qui s'avèrent être opposées selon leur tension, ou plutôt leur appartenance à une catégorie *fortis* et l'autre, *lenis* (Kohler, 1984). Ainsi, il n'y a pas de correspondance entre les occlusives du français et celles de l'alsacien, ce qui est un vecteur de ce que les francophones perçoivent comme étant un accent alsacien. Notons que cette problématique s'applique également aux fricatives, telles que `/f/` et `/v/` par exemple. Bien entendu, les systèmes graphiques alsaciens utilisent les mêmes symboles qu'en français, mais il n'est pas simple de faire un choix entre les deux graphèmes, aucune des prononciations françaises n'étant correcte. Ces phonèmes posent le problème de consonnes le plus massif : l'analyse des disparités entre les lexiques montre une hésitation fréquente sur l'usage des graphèmes `p,t,k` et `b,d,g`. En effet, ces consonnes sont très souvent utilisées les unes à la place des autres, dans 278 cas au total, par opposition à seulement 10 cas de modifications des fricatives, telles que `/f–v/` ou encore `/s–z/`. La paire apico-alvéodentale, graphiée `t,d`, occupe la neuvième place des modifications les plus fréquentes, après les nombreux changements de voyelles. Elle représente à elle seule 166 modifications, tant en position initiale (ex. *dànze – tànze*, danser) qu'en intervocalique (ex. *Vàder – Vàter*, père) ou qu'en finale de mot (ex. *G'hàlt – Gald*, argent). Un problème certain est soulevé par ce phénomène : il existe de nombreuses paires minimales opposées uniquement par la consonne occlusive, comme *Pump* (pompe) et *Bumb* (bombe), ou *Gàss* (ruelle) et *Kàss* (caisse). Dans ces cas, il serait absolument nécessaire d'opérer une distinction entre les graphèmes utilisés, mais l'étude des lexiques montre que les hésitations sont nombreuses lors de la mise à l'écrit de ces phonèmes spécifiquement. Ainsi, les consonnes elles-mêmes ne sont parfois pas fiables, ce qui est compensé dans l'algorithme *Double Metaphone* par la neutralisation de la différence entre ces consonnes (voir Table 2 : le 'd' dans "Rüejd̲àà" et le 't' dans "Rüaijt̲àág" sont indifféremment transcrits par 'T' dans les clés metaphone). L'utilisation de ressources globales permettant la comparaison, par exemple, avec les graphies choisies en allemand, pourrait contribuer à stabiliser des formes écrites qui respecteraient toujours l'opposition *fortis-lenis*. Cette stabilisation est un enjeu d'importance au vu de l'existence de paires minimales dont l'opposition repose uniquement sur ces consonnes. Il s'agirait donc de tendre à normaliser l'usage de ces graphèmes, à travers une comparaison des formes existantes dans divers dictionnaires, afin de permettre de clarifier les choix à faire dans les ressources futures.

6 Conclusion et perspectives

L'absence de convention orthographique est un problème pour de nombreuses langues peu dotées en ressources linguistiques, ce qui complique encore davantage l'acquisition de ressources lexicales. Nous avons proposé des solutions qui utilisent des ressources disponibles facilement (lexiques bilingues, réseau sémantique BabelNet) et des outils simples à développer et à adapter pour de nouvelles langues (*Double Metaphone*). Les ressources obtenues sont mises en correspondance avec BabelNet, qui ajoute une couche sémantique et donne accès à différents types d'informations supplémentaires : définitions et gloses, traductions dans d'autres langues, des images, etc.

La méthode proposée pour l'alignement des lexiques vise la précision plutôt que le rappel et peut être utilisée pour construire facilement et rapidement des ressources lexicales multilingues avec une intervention humaine limitée pour la correction. Elle pourrait en principe être appliquée à de nombreuses autres langues, car elle nécessite peu de ressources. Son originalité est qu'elle ne cible pas la normalisation, mais vise plutôt à regrouper les variantes graphiques et à les relier à des entrées dans une ressource multilingue. De cette manière, les ressources du *Linguistic Linked Open Data* peuvent être étendues avec des lexicalisations de langues peu dotées et aider à construire et enrichir les ressources pour ces langues. L'alignement des lexiques a mis en évidence les difficultés de scripturalisation des dialectes alsaciens, en lien avec des faits phonétiques avérés. Les consonnes en particulier posent des difficultés résultant en de nombreuses hésitations qui peuvent compliquer l'alignement.

Dans l'avenir, nous prévoyons de fournir le lexique aligné dans un format standard. Nous souhaitons également améliorer les alignements grâce à un meilleur algorithme *Double Metaphone*, une analyse appropriée des formes de mots composés et l'utilisation des ressources supplémentaires avec une meilleure couverture. Enfin, le lexique obtenu pourra être utilisé dans diverses applications, par exemple l'étiquetage morphosyntaxique (Bernhard & Ligozat, 2014).

Remerciements Nous remercions l'OLCA, André Nisslé et Paul Adolf pour nous avoir donné accès à leurs ressources. Les travaux décrits dans cet article ont bénéficié du soutien de l'ANR (projet RESTAURE - convention ANR-14-CE24-0003-01) et du conseil scientifique de l'Université de Strasbourg (projet COPAL).

Références

ADOLF P. (2006). *Dictionnaire comparatif multilingue : français-allemand-alsacien-anglais*. Strasbourg, France : Midgard.

ALMEIDA J. J., SANTOS A. & SIMÕES A. (2010). Bigorna – A Toolkit for Orthography Migration Challenges. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta.

BANEA C. & MIHALCEA R. (2011). Word sense disambiguation with multilingual features. In *Proceedings of the Ninth International Conference on Computational Semantics*, p. 25–34.

BAUER L. (2003). *Introducing Linguistic Morphology*. Georgetown University Press. 2nd edition.

BERNHARD D. & LIGOZAT A.-L. (2014). 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.

BOTHOREL-WITZ A. & PÉTURSSON M. (1972). La nature des traits de tension, de sonorité et d'aspiration dans le système des occlusives de l'allemand et de l'islandais. *Travaux de L'Institut de Phonétique de Strasbourg*, (4).

DASIGI P. & DIAB M. (2011). CODACT : Towards Identifying Orthographic Variants in Dialectal Arabic. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, p. 318–326, Chiang Mai, Thailand.

ERHART P. (2012). *Les dialectes dans les médias : quelle image de l'Alsace véhiculent-ils dans les émissions de la télévision régionale ?* Thèse de doctorat, Université de Strasbourg.

GUREVYCH I., ECKLE-KOHLER J., HARTMANN S., MATUSCHEK M., MEYER C. M. & WIRTH C. (2012). UBY–A Large-Scale Unified Lexical-Semantic Resource Based on LMF. In *Proceedings of EACL*, p. 580–590, Avignon, France.

HANOKA V. & SAGOT B. (2012). Wordnet creation and extension made simple : A multilingual lexicon-based approach using wiki resources. In *Proceedings of the 8th international conference on Language Resources and Evaluation (LREC 2012)*.

HUCK D., BOTHOREL-WITZ A. & GEIGER-JAILLET A. (2007). L'Alsace et ses langues. Éléments de description d'une situation sociolinguistique en zone frontalière. *Aspects of Multilingualism in European Border Regions : Insights and Views from Alsace, Eastern Macedonia and Thrace, the Lublin Voivodeship and South Tyrol*, p. 13–100.

HUDLETT A. & GROUPE D'ETUDES ET DE RECHERCHES INTERDISCIPLINAIRES SUR LE PLURILINGUISME EN ALSACE ET EN EUROPE (2003). *Charte de la graphie harmonisée des parlers alsaciens : système graphique GRAPHAL - GERIPA*. Mulhouse, France : Centre de Recherche sur l'Europe littéraire (C.R.E.L.).

HULDEN M., ALEGRIA I., ETXEBERRIA I. & MARITXALAR M. (2011). Learning word-level dialectal variation as phonological replacement rules using a limited parallel corpus. In *Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*, p. 39–48, Edinburgh, Scotland.

INKPEN D., FRUNZA O. & KONDRAK G. (2005). Automatic identification of cognates and false friends in French and English. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, p. 251–257.

KLEIBER G. & RIEGEL M. (1998). Grammaticalisation et auxiliaire modal : L'énigme de duen en Alsacien. In *Travaux de linguistique*, volume 36, p. 161–173, Bruxelles, Belgique : Rijksuniversiteit van Gent.

KOHLER K. (1984). Phonetic explanation in phonology : the feature fortis/lenis. *Phonetica*, **41**, 150–174.

LAFOURCADE M. (2007). Making people play for Lexical Acquisition. In *Proceedings of SNLP 2007*, Pattaya, Thailande.

LEIXA J., MAPELLI V. & CHOUKRI K. (2014). *Inventaire des ressources linguistiques des langues de France*. Rapport ELDA/DGLFLF-2013A, ELDA, Paris.

MATZEN R. (1985). Les emprunts du dialecte alsacien au français. In *Le français en Alsace : Actes du colloque de Mulhouse (17-19 novembre 1983)*, Bulletin de la Faculté des lettres de Mulhouse, Mulhouse : Paris : Champion, Genève : Slatkine.

MOSQUERA A., LLORET E. & MOREDA P. (2012). Towards Facilitating the Accessibility of Web 2.0 Texts through Text Normalisation. In *Proceedings of the Workshop on Natural Language Processing for Improving Textual Accessibility (NLP4ITA)*.

NAVIGLI R. & PONZETTO S. P. (2012). BabelNet : The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, **193**, 217–250.

NIEMANN E. & GUREVYCH I. (2011). The people's web meets linguistic knowledge : Automatic sense alignment of Wikipedia and WordNet. In *Proceedings of the 9th International Conference on Computational Semantics (IWCS)*, p. 205–214.

OLCA / EDINSTITUT (2012). Etude sur le dialecte alsacien. En ligne : `https://www.olcalsace.org/sites/default/files/documents/etude_linguistique_olca_edinstitut.pdf`.

PHILLIPS L. (2000). The Double Metaphone Search Algorithm. *C/C++ Users Journal*.

PIPE K. (2014). *Accent Levelling in the Regional French of Alsace*. Thèse de doctorat, University of Exeter.

PORTA J., SANCHO J.-L. & GÓMEZ J. (2013). Edit Transducers for Spelling Variation in Old Spanish. In *Proceedings of the Workshop on Computational Historical Linguistics at NoDaLiDa 2013*, volume 87, p. 70–79.

SAGOT B. & FIŠER D. (2008). Construction d'un wordnet libre du français à partir de ressources multilingues. In *Actes de TALN 2008-Traitement Automatique des Langues Naturelles*.

SALLOUM W. & HABASH N. (2011). Dialectal to standard Arabic paraphrasing to improve Arabic-English statistical machine translation. In *Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*, p. 10–21.

SCHERRER Y. (2008). Transducteurs à fenêtre glissante pour l'induction lexicale. In *Actes de RECITAL 2008*, Avignon.

SCHMID H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*, p. 44–49.

STEIBLÉ L. (2014). *Le contrôle temporel des consonnes occlusives de l'alsacien et du français parlé en Alsace*. Thèse de doctorat, Université de Strasbourg.

TANAKA K. & UMEMURA K. (1994). Construction of a bilingual dictionary intermediated by a third language. In *Proceedings of the 15th conference on Computational linguistics-Volume 1*, p. 297–303.

WOEHRLING C. & BOULA DE MAREÜIL P. (2005). Identification d'accents régionaux en français : perception et catégorisation. *Bulletin PFC 6*, p. 89–102.

ZEIDLER E. & CRÉVENAT-WERNER D. (2008). *Orthographe alsacienne : bien écrire l'alsacien de Wissembourg à Ferrette*. Colmar, France : J. Do Bentzinger.

PICARTEXT : Une ressource informatisée pour la langue picarde

Jean-Michel Eloy[1], Fanny Martin[1], Christophe Rey[1]

(1) LESCLAP (CERCLL-EA 4283), Université de Picardie Jules Verne, Amiens
jean-michel.eloy@u-picardie.fr, fanny.martin@u-picardie.fr, christophe.rey@u-picardie.fr

Résumé.

Picartext est une base de données textuelles, construite depuis près de 10 ans à l'Université de Picardie à Amiens. Elle présente des caractéristiques de premier intérêt pour la recherche sur les traitements automatiques. La langue picarde, d'une vitalité non négligeable, dispose d'une littérature assez abondante et de très nombreux dictionnaires et glossaires. Mais elle ne possède pas de standard, ni linguistique, ni graphique. La langue est donc très variante. La base de données, de nature littéraire, d'environ 5 millions d'occurrences, est accessible en ligne au moyen d'un outil d'interrogation paramétrable : non seulement il permet la restriction du corpus de travail (lieux, dates, genres), mais il permet une recherche tenant compte d'équivalences phonétiques et d'équivalences dialectales. Il est ouvert à des évolutions en termes de balisage, en particulier dans le cadre d'un projet ANR portant sur trois langues régionales simultanément (picard, alsacien, occitan).

Abstract.

PICARTEXT : a computerized resource for picard

Picartext is a textual database, built up since about 10 years in Picardy University in Amiens. Some of its characteristics make it very interesting for research on natural languages processing. Picard language, of a not insignificant vitality, has a rather plentiful literature, and very numerous dictionaries and glossaries. But it does not possess standard, either linguistics, or graphic. The language is thus very variant. The database, of literary nature, counts about 5 million token, is reachable on-line, with a customizable tool of interrogation : not only it allows the limitation of the working corpus (places, dates, genres), but he allows a search taking into account phonetic equivalences and dialectal equivalences. It is opened to evolutions in terms of tagging, in particular within the framework of an ANR project concerning three regional languages simultaneously (picard, alsatian, occitan).

Mots-clés :

picard, non standardisation, variation dialectale, variation graphique, numérisation, balisage, équivalences

Keywords:

picard language, non standardisation, dialectal variation, graphical variation, digitisation, tagging, equivalences

1 La langue picarde

1.1 Données générales sur le picard, langue de France

La langue picarde se trouve sur un territoire vaste, mais divisé puisqu'elle s'étend sur deux régions françaises et une région de Belgique. En dépit de cette situation géographique, son originalité et son unité sont assez fortes. La carte reproduite ci-dessous permet de se figurer son domaine linguistique :

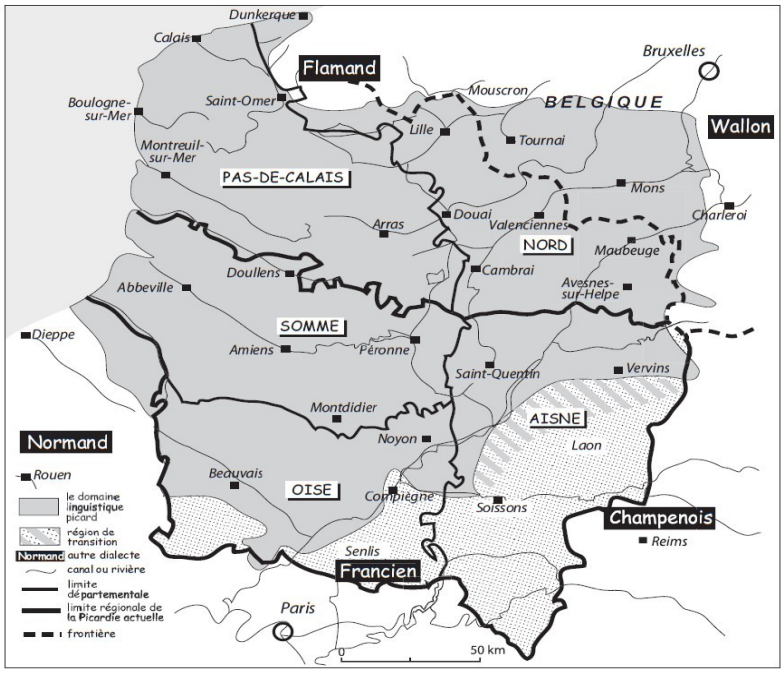


FIGURE 1: L'aire linguistique picarde, carte établie par René Debrie selon les informations de Raymond Dubois, réalisée par Joëlle Désiré pour l'Atlas de Picardie (Amiens, Université de Picardie Jules Verne)

La région connaît un assez fort investissement identitaire, mais il existe une coupure entre *picard* et *patois, chtimi, rouchi*[1]. Le statut de la langue, encore discuté aujourd'hui, est marqué par une reconnaissance incomplète par les pouvoirs publics[2]. En effet, alors que cette reconnaissance est moyenne en région Picardie – dans la mesure où le picard peut s'appuyer sur l'existence de discours de promotion, notamment financés par des fonds publics – elle est en revanche très faible en région Nord-Pas-de-Calais, comme au niveau national, et ce malgré le rapport Cerquiglini de 1999.

La langue picarde appartient au groupe gallo-roman, dans une relation de proximité dynamique « langues collatérales » (Eloy 2004) avec le français. C'est une « variété basse de diglossie » assez typique – et qui fait l'objet actuellement d'efforts de « retroussement de la diglossie » (Lafont 1984).

1.2 Un développement sans standardisation

L'historicité de la littérature en langue picarde est considérable. Les pratiques littéraires dans cette variété s'étendent en effet depuis le Moyen-Âge (dans le cadre d'une langue d'oïl tolérante), se renforcent au XVII^{e} siècle, et connaissent ensuite une croissance constante au XIX^{e} siècle. Les publications actuelles se comptent quant à elles par dizaines chaque année.
En dépit de cette histoire littéraire déjà longue et riche, le picard est caractérisé par un développement sans standardisation, qui semblerait se faire au bénéfice de « pôles de pratiques » (Forlot & Martin, 2014 ; Martin, 2015; Martin & Forlot, à paraître).

[1] Bien qu'il s'agisse de la même langue, il existe néanmoins des variétés. Ces appellations correspondent à des nominations différentes sur le continuum géographique.
[2] Mais également par les habitants du domaine linguistique, locuteurs ou non locuteurs du picard.

Considérons par exemple ce qui se passe pour la production lexicographique dans cette langue. Le picard jouit d'une richesse lexicographique sans égale par rapport aux autres langues régionales de France puisque l'on peut recenser plusieurs centaines de titres proposés depuis le XVIIIe siècle jusqu'à nos jours. Existant en l'absence d'une standardisation, cette richesse s'explique à la fois par la tradition dialectologique, une forme d'aménagement du statut (de nombreux lexiques et glossaires se trouvent ainsi insérés en fin de volumes de récits) et une « grammatisation militante » : dictionnaires français-picard, dictionnaires en ligne.
Bien sûr, cette situation de non standardisation est liée au statut de la langue qui ne bénéficie d'aucun enseignement (Forlot & Martin, 2015), d'aucun usage officiel, ni de politique linguistique véritable. Et même actuellement, il n'y a aucune demande de la part des acteurs du monde picardisant pour faire évoluer cette configuration.

Intéressons-nous ensuite à la question de l'orthographe en picard, dimension qui confirme elle aussi l'absence d'une standardisation, malgré de nombreux débats antérieurs.
L'orthographe picarde voit la coexistence de quelques systèmes et d'une marge anarchique. Citons en exemple la séquence « c'était » que l'on peut ainsi retrouver sous les formes *ch'étoué, ch'étouait, ch'étouis, ch'étouo, ch'étwo, ch'étouot, ch'étwot,* et même *ché toué*, etc. Cette multiplicité de formes traduit certes des phénomènes de variation dialectale à l'intérieur de l'espace linguistique picard, mais elle atteste également une tendance forte à la variabilité graphique et à des problèmes de segmentation possiblement imputables à l'absence de standard, phénomènes particulièrement manifestes dans la base PICARTEXT.
Un second exemple illustrant la grande richesse orthographique du picard peut être donné à travers l'évocation du *Dictionnaire général français-picard* de Jean-Marie Braillon, ouvrage de 2001 dans lequel sont en effet recensées pas moins de 18 graphies différentes pour le mot « aiguille ».

L'un des points les plus saillants de la variabilité orthographique du picard (Dawson, 2002) concerne sans doute l'utilisation des apostrophes qui permet de dégager des séquences aussi distinctes que les suivantes: *chol vaque, chov vaque, chov'vaque, cho'v vaque/ch'timi, chtimi/pèmes tère, pèm'terre*, etc.

En pratique, on retiendra qu'il existe quatre grands types de graphies pour la langue picarde - dont trois constituent des substandards -, à savoir l'orthographe proposée par Vasseur (Vasseur, 1968), celle proposée par Feller et Carton (Carton, 2001), celle livrée par Braillon (Braillon, 1991), et enfin les cacographies. Soulignons toutefois que de récents succès de librairie, dont une traduction des albums d'Astérix tirée à 130 000 exemplaires, ont vu la mise en évidence de substandards.
En bref, malgré l'historicité, et la nette unité, la caractéristique centrale en picard est la variation maximale : une variation en liberté, qui est aussi une forme de contrainte dans son expansion et se traduit par une « double insécurité » (Martin, 2015) chez les locuteurs et les néo-locuteurs.

2 PICARTEXT

Nous livrons ci-dessous un aperçu synthétique de la base de données PICARTEXT conçue au LESCLAP, en esquissant successivement une description de la nature du corpus textuel mis en place et en apportant ensuite des éléments d'information concernant les modalités techniques d'informatisation et d'exploitation de celle-ci.

2.1 Le corpus Picartext

La base de données[3] a été conçue en fonction de problématiques linguistiques, et non pas purement pour le Traitement Automatique des Langues. C'est un travail qui se situe dans la continuité du Centre d'Etudes Picardes (CEP) - hébergé désormais par le laboratoire *Linguistique Et Sociolinguistique : Contacts, Lexique, Appropriations, Politiques (LESCLAP)*. L'équipe picarde travaille par ailleurs sur les problématiques de langues minorées, de langues proches, et de politiques linguistiques.

Nous avions dans l'équipe une connaissance d'utilisateurs de Frantext, et une expérience de travaux de rétroconversion de dictionnaires anciens. Des études de faisabilité ont été réalisées par des étudiants de Lille inscrits dans un Master de lexicologie et lexicographie. En tant que linguistes, nous avions une bonne connaissance de la langue et de la littérature

[3]La base PICARTEXT est constituée de textes écrits partiellement ou totalement en picard, issus de l'ensemble du domaine linguistique picard, et composés depuis le XVIIIe siècle jusqu'à nos jours. Son objectif est d'offrir à la communauté des chercheurs, ainsi qu'à un public averti, une ressource linguistique à partir de laquelle il sera possible d'envisager toutes sortes d'exploitations :
- exploration de la langue (lexique, morphosyntaxe, phraséologie...)
- étude des évolutions diachroniques
- étude de la variation et de la cohésion dialectales et du processus de koinèisation

picardes. Deux informaticiens-linguistes ont ensuite, grâce à des contrats postdoctoraux, réalisé la base. (Eloy-Rey-Dawson, 2011)

Le contenu de la base est un corpus de littérature d'environ cinq millions de mots. Il est difficile de dire ce que représente la totalité des publications en picard : 80 millions d'occurrences ? 150 millions ? Plus ? Notons que la question est insoluble aussi, à une autre échelle, en français.
Le problème de la sélection des textes, malgré un effort de rationalisation, a été réglé empiriquement (sur la base de « noms bien connus »). Il semble d'ailleurs que le grand modèle qu'est Frantext[4] n'ait pas pu procéder autrement. La difficulté augmente avec le développement de la littérature, donc elle est plus grande pour le XX[e] siècle que pour le XVIII[e] siècle.
La diversité est très grande, saisie en termes de lieux, de dates et de genres. Les genres, par exemple, jouent diversement sur la qualité de la langue : imitation du parler, imitation du français, lyrisme, vulgarité, etc.
La notion d'œuvre (ou même de texte) pose quelques problèmes. Le premier est celui de l'alternance, car il existe des textes alternants, ou bilingues, ou simplement des préfaces en français. Pour dégager un corpus seulement en picard, nous avons donc été dans l'obligation de sous-délimiter au sein des œuvres. Les scholies de théâtre, par exemple, sont parfois en français, parfois en picard, ce qui rend les choix délicats concernant ce type de « paratexte ». Plus embarrassante encore est la langue intermédiaire de certains textes, constituant une sorte de continuum souvent nommé « patois » (et non langue). Bref, la notion d'œuvre en picard implique des interventions sélectives, en partie arbitraires car dans cette culture les langues vivent ensemble et nous voulons les séparer.

2.2 Le processus de numérisation

La constitution de la base de données PICARTEXT s'est appuyée sur la collecte de quelques textes déjà numérisés, mais a surtout consisté à intégrer des textes majoritairement édités sur papier. Nous avons pour cela mis au point un itinéraire du texte édité, qui passe par plusieurs étapes. Les documents sélectionnés ont d'abord été photocopiés avant d'être traités par un logiciel de reconnaissance automatique de caractères. La qualité du rendu a tout de même nécessité une importante phase de double relecture des textes informatisés avant leur livraison sous un format .TXT considéré comme exploitable pour la suite des traitements informatiques et notamment la phase d'application d'un balisage logique au format XML. Le schéma reproduit ci-dessous illustre le parcours des textes intégrés dans PICARTEXT :

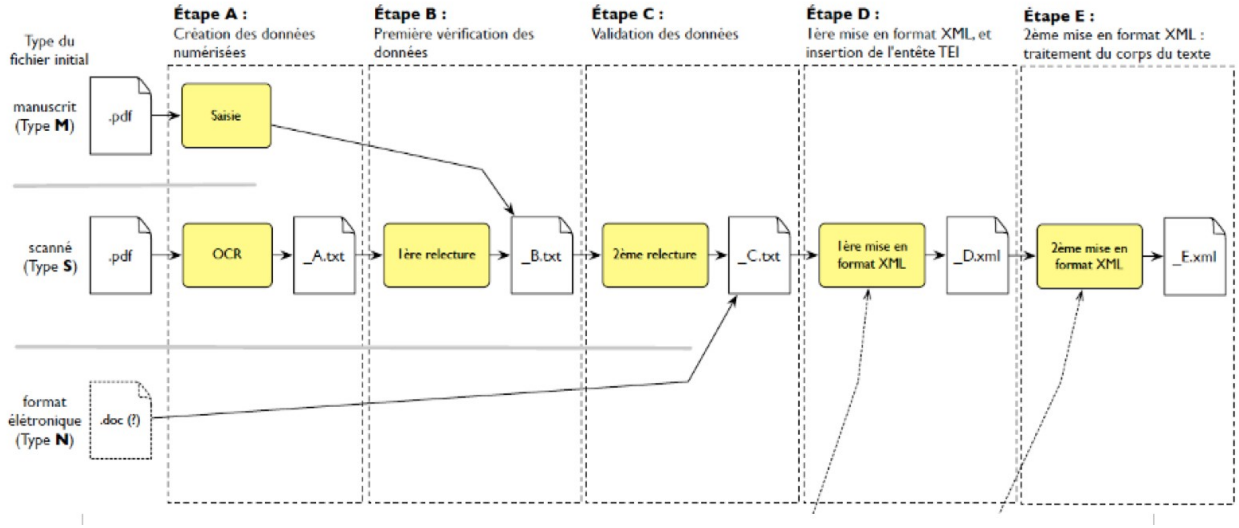


FIGURE 2 : Parcours des textes dans PICARTEXT

2.3 Description du module d'interrogation

La base PICARTEXT, désormais accessible en ligne par le biais d'un portail internet dédié[5], bénéficie d'un module d'interrogation répondant à des finalités du grand public et des chercheurs.

4 http://www.frantext.fr/
5 https://www.u-picardie.fr/LESCLaP/PICARTEXT/Public/

Sans être obligatoire, l'interrogation de la base PICARTEXT permet de déterminer ou de restreindre un corpus de travail, selon plusieurs paramètres : lieux, dates, genres. Puis vient le choix d'une méthode d'interrogation.
Outre la recherche par « chaînes de caractères » et celle par « expressions rationnelles », on peut procéder aussi par deux méthodes plus élaborées, appuyée sur Dawson (2006), utilisant l'approche théorique de Mc Carthy & Prince (1995) :
- par « correspondance phonétique » : le mot est d'abord converti en sa représentation phonétique à l'aide d'un phonétiseur[6]. C'est cette représentation phonétique qui est recherchée, ce qui permet de ne pas tenir compte de l'orthographe des auteurs.
- par « correspondance dialectale » : le mot est converti en une forme abstraite (lemme dialectal) qui neutralise la variation dialectale du picard. Ceci permet de retrouver le mot sous diverses formes dialectales.
Le site donne des exemples de ces différentes modalités de requête.

La publication des résultats se fait actuellement par le biais de concordances récupérables sous la forme de données tabulaires ou au simple format texte.

2.4 Une base configurée pour être développée : la voie incontournable du balisage XML

Il est toujours à craindre que l'investissement humain et financier ne permette pas immédiatement de pousser les travaux aussi loin qu'on le voudrait. Notre stratégie a donc consisté à assurer d'abord la structuration du matériau linguistique - qui reste toujours enrichissable et continue d'ailleurs à s'enrichir - et les outils d'interrogation à disposition du public spécialiste ou non. Les standards TEI[7] et XML[8] ont ainsi été retenus, notamment dans une perspective de pérennisation de la ressource.
Le projet PICARTEXT s'est ainsi intégré à la galaxie des très nombreuses ressources pouvant bénéficier des travaux de balisage dans le langage XML. L'étape à laquelle nous arrivons est précisément celle de la mise en place d'un protocole de balisage logique XML conforme à la TEI qui puisse tenir compte de la grande diversité des types de documents (poésies, recueils de prose, pièces de théâtre, chansons, dictionnaires, etc.) constituant la base.

L'extension qualitative de la base nous amènera à y intégrer davantage de textes anciens. Nous avons aussi commencé à nous interroger sur la possibilité d'y ajouter des transcriptions d'oral. Ces différentes perspectives d'évolution nous amèneront nécessairement à parfaire le module d'interrogation lui-même et à le rendre capable de gérer au mieux l'accroissement des possibilités de recherche.

3 Les paris du projet RESTAURE

Depuis peu, le LESCLAP est l'un des partenaires scientifiques du projet de recherche ANR RESTAURE. Le LESCLAP avait pris l'initiative de se confronter à d'autres langues de situations similaires lors d'une journée d'étude qui préfigurait le projet RESTAURE, concrétisé grâce à des collègues travaillant sur l'alsacien et d'autres sur l'occitan.

Visant à « fournir des ressources informatiques et des outils de traitement automatique pour trois langues régionales de France : alsacien, occitan et picard », le projet RESTAURE repose sur le développement « de nouveaux modèles adaptés aux langues disposant de peu de ressources et peu standardisées ». Les procédures d'enrichissement des ressources, comme les traitements, devront prendre en compte la non standardisation, la variation linguistique et graphique, voire les faibles moyens financiers. *In fine*, l'objectif est de disposer de ressources textuelles, de corpus annotés et à partir de là, de lexiques morphologiques étendus, tout d'abord dans les trois langues envisagées, avec extension des méthodes à d'autres langues « peu dotées ». L'utilité sociale concerne donc les communautés linguistiques autant que les milieux de la recherche.

Le LESCLAP, fort de l'expérience de PICARTEXT, voit dans ce projet une opportunité de prolonger les acquis de cette ressource. Un des objectifs de ce programme, en 2008, était d'intéresser des informaticiens aux traitements de cette langue, et nous notons avec satisfaction, ainsi qu'en témoigne le nombre de visites du portail internet l'hébergeant, que la base jouit d'un intérêt véritable.

---

[6] Le phonétiseur utilisé dans le module expérimental de recherche dans le corpus Picartext est issu du système TTS-French développé par David Haubensack, sur la base des travaux de Thierry Dutoit, dans le cadre du projet MBROLA de la Faculté Polytechnique de Mons (Belgique). Références :
- Dutoit, T., V. Pagel, N. Pierret, F. Bataille, O. van der Vreken, 1996. "The MBROLA Project: Towards a Set of High-Quality Speech Synthesizers Free of Use for Non-Commercial Purposes" Proc. ICSLP'96, Philadelphia, vol. 3, pp. 1393-1396.
- Dutoit, Thierry. 1997. *An Introduction to Text-To-Speech Synthesis*. Kluwer Academic Publishers, Dordrecht Hardbound.

[7] TEI Consortium, <http://www.tei-c.org/Guidelines/P5/>.

[8] World Wide Web Consortium, <http://www.w3.org/TR/xml/>

Les attentes et les objectifs du projet RESTAURE, en ce qui concerne plus particulièrement la langue picarde, sont nombreux. Ce projet est par exemple une occasion d'étendre le nombre de textes en langue picarde mis à disposition sous forme numérique. Grâce à la mise en place d'un processus de numérisation améliorant fortement les réalisations de PICARTEXT, nous serons en mesure de détenir un dispositif méthodique et pérenne de constitution de nouvelles données.

Une des particularités très fortes de la langue picarde réside dans l'extrême variation graphique à laquelle elle est sujette. Le projet RESTAURE visant à mettre au point des outils de maîtrise de cette variation par la proposition de règles d'équivalence, nous nous interrogeons sur la possibilité de trouver une ou plusieurs solutions véritablement satisfaisantes. Ces règles pourront-elles d'ailleurs être éventuellement transversales aux trois langues du projet ou resteront-elles propres à chaque langue ? Voilà l'un des enjeux forts de RESTAURE.

Une caractéristique commune de l'alsacien, de l'occitan et du picard, est qu'il n'existe pour aucune de ces langues de dictionnaire faisant office de « standard ». Les outils automatiques de désambiguïsation grammaticale et lexicale envisagés dans le cadre du projet devront donc s'appuyer sur des procédures d'étiquetage morphosyntaxique particulièrement robustes. Il s'agit là d'une perspective particulièrement stimulante qui pourra notamment s'appuyer sur l'utilisation des dictionnaires de langues proches, par exemple allemand standard et français standard. On notera donc que le bénéfice ira aux méthodes, donc à toutes les langues, y compris celles déjà « bien dotées ».

Conclusion

Arrivé au terme de son financement, le projet PICARTEXT s'impose comme une ressource linguistique de tout premier ordre pour le picard. Mettant en évidence une langue qui se construit en liberté dans la variation et en absence de standardisation, ce projet doit encore être valorisé auprès du grand public pour atteindre les objectifs de sa création.
En ce qui concerne la communauté scientifique, PICARTEXT comble déjà toutes nos attentes puisqu'il offre non seulement des perspectives de recherche conséquentes, mais fait aussi entrer le picard dans le cercle restreint des langues régionales de France disposant, grâce à l'informatique et au Traitement Automatique des Langues, d'une visibilité accrue. L'intégration récente de l'équipe LESCLAP au projet RESTAURE constitue, selon nous, une illustration de l'intérêt de cette langue pour les recherches sur les autres variétés régionales et les langues plus richement dotées. Bénéficiant des initiatives audacieuses de ce projet, la base PICARTEXT devrait, sans nul doute, connaître dans les années futures des avancées considérables.

Remerciements Les travaux décrits dans cet article ont bénéficié du soutien de l'ANR (projet RESTAURE - convention ANR-14-CE24-0003-01) et du Conseil régional de Picardie.

Références

BRAILLON J.-M. (2001). *Dictionnaire général français-picard*, Tome I, éditions F.I.P.Q.

BRAILLON J.-M. (1991). *La graphie FIPQ du picard*, Lemé.

CARTON F. (2001). « Orthographe picarde Feller-Carton », in Linguistique Picarde, décembre.

CERQUIGLINI B. (1999). *Les Langues de la France, Rapport au Ministre de l'Éducation nationale, de la Recherche et de la Technologie*.

DAWSON A., ELOY J.-M., REY C. (2011 – non publié). Vue perspective sur le français à partir d'une base de données textuelles en domaine d'oïl, *Colloque annuel de l'Association for French Language Studies*, 8-10 septembre 2011, Nancy.

DAWSON A. (2006). *Variation phonologique et cohésion dialectale en picard. Vers une Théorie des Correspondances Dialectales*, Thèse de doctorat sous la direction de Marc PLÉNAT. 340 pages.

DAWSON A. (2002). « Le picard, langue polynomique, langue polygraphique ? », in D. Caubet, S. Chaker, J. Sibille (éd.*), La codification des langues de France*, L'Harmattan, Paris.

DUTOIT T., PAGEL V., PIERRET N., BATAILLE F., VAN DER VREKEN O. (1996). « The MBROLA Project: Towards a Set of High-Quality Speech Synthesizers Free of Use for Non-Commercial Purposes », Proc. ICSLP'96, Philadelphia, vol. 3, 1393-1396.

DUTOIT T. (1997). *An Introduction to Text-To-Speech Synthesis*. Kluwer Academic Publishers, Dordrecht Hardbound .

ELOY J.-M., REY C. (2012 – non publié). Une base de données lexicale en picard : la base PICARTEXT, *Journée d'étude du LESCLAP : "Bases de données informatisées dans les "petites langues""*, Amiens, 07 décembre 2012.

ELOY J.-M. (2004). *Des langues collatérales*, Paris, L'Harmattan.

FORLOT G., MARTIN F. (2014). « Entre invisibilité et (auto)occultation. Les paradoxes des pratiques langagières minoritaires en Picardie », in K. Djordjevic (éd.), Les minorités invisibles : diversité et complexité (ethno)sociolinguistiques, Éditions Lambert-Lucas, Limoges, pp. 77-87.

FORLOT G., MARTIN F. (2015). « Le picard à l'épreuve du terrain scolaire aujourd'hui », Communication dans le cadre du Congrès international de sociolinguistique, Grenoble, 2015.

LAFONT R. (1984). Pour retrousser la diglossie. *Lengas*, 15, 5-36.

MARTIN F., FORLOT G. (à paraître). « Hétérogénéité linguistique et poids des idéologies sur les pratiques linguistiques en Picardie », in A. Boudreau et L. Arrighi, *La construction discursive du locuteur francophone en milieu minoritaire. Problématiques, méthodes et enjeux*, Presses de l'Université Laval, Ste Foy (Québec).

MARTIN F. (2015). *Espaces et lieux de la langue en Picardie au XXIème siècle. Approche complexe de la structuration des répertoires linguistiques en situations ordinaires. Enquête en Picardie*, Thèse de doctorat, Université de Picardie Jules Verne, Amiens.

MCCARTHY J., PRINCE A. (1995). Faithfullness and Reduplicative Identity, *in* J. Beckman, L. Walsh Dickey, S. Uurbanczyk (éd.), *Papers in Optimality Theory*, U. of Massachusetts Occasional Papers in Linguistics 18, Amherst, Mass. : Graduate Linguistic Student Association, 249-384.

VASSEUR G. (1968). « L'orthographe picarde, principes généraux et règles pratiques établis par les Picardisants du Ponthieu et du Vimeu », *Linguistique Picarde*, décembre 1968 (graphie analogique).

Akenou-Breizh, un projet de plate-forme valorisant des ressources et outils informatiques et linguistiques pour le breton

Annie Foret[1] Valérie Bellynck[2] Christian Boitet[2]
(1) IRISA & Université Rennes 1, Campus de Beaulieu, F-35042 Rennes cedex
(2) LIG-GETALP, UdG &CNRS, Campus, 41 rue des Mathématiques, GRENOBLE cedex 9
Annie.Foret@irisa.fr, Valerie.Bellynck@imag.fr, Christian.Boitet@imag.fr

Résumé. Nous présentons un nouveau projet, *Akenou-Breizh*, qui vise (1) à mettre en place une plate-forme permettant d'étudier les influences d'une *langue d'héritage*, comme le breton, sur une *langue d'usage*, comme le français, et (2) à mettre à disposition de tous les intéressés des outils s'intégrant au "Web sémantique et multilingue", et proposant des accès proactifs aux connaissances sur le breton ainsi qu'une visualisation directe des correspondances sous-phrastiques dans des présentations bilingues alignées. Nous nous proposons non seulement d'utiliser les nombreuses ressources disponibles librement, en particulier celles de l'OPLB [1] et du projet APERTIUM, mais aussi d'en créer de nouvelles, comme des corpus bilingues alignés de bonne qualité, en utilisant le "Web collaboratif", et de construire sur le site dédié `lingwarium.org` des modules linguistiques améliorant ou étendant ceux qui existent, par exemple un analyseur-générateur morphologique. Nous décrivons aussi une expérience réalisée à partir d'un lexique réduit pour le breton, qui montre comment on peut enrichir un dictionnaire classique, en le reliant à un treillis de thèmes et à un système de gestion de contexte (ici CAMELIS), de façon à ce qu'on puisse l'interroger (par facettes sémantiques) et comparer différentes ressources.

Abstract.

***Akenou-Breizh*, a platform project to develop computational and linguistic resources and tools for breton**

We present a new project, *Akenou-Breizh*, that aims to (1) put in place a platform allowing to study the influences of an *heritage language*, such as Breton, on a *usage language*, such as French, and (2) to make available, to all interested persons, tools well integrated in the "semantic and multilingual web" and proposing proactive access to various kinds of knowledge concerning Breton, as well as direct visualisation of infrasentential correspondences in aligned bilingual presentations. We plan not only to use the numerous freely available resources, in particular those of OPLB and of the APERTIUM project, but also to create new ones, such as good quality bilingual aligned corpora, thereby using the "collaborative web", and to build on the dedicated lingwarium.org web site linguistic modules improving on or extending those that exist, for example a morphological analyzer-generator. We also describe an experiment set up starting from a reduced lexicon for Breton, that shows how it is possible to enrich a classical dictionary, by linking it to a lattice of topics and to a context management system (here CAMELIS), in such a way one can query it (along semantic facets) and compare different resources.

Mots-clés : breton, langue d'héritage, langue d'usage, outils et ressources, études contrastives.

Keywords: Breton, heritage language, usage language, tools and resources, contrastive studies.

1 Problématique et enjeux

On cite souvent Claude Hagège : *les langues sont les drapeaux des identités nationales*. C'est très vrai, et l'on voit des variantes de langue être érigées en langues distinctes pour cette raison, par exemple l'hindoustani en hindi et ourdou, ou le serbo-croate en serbe, croate et bosniaque. L'article 2 de la Constitution Française, dans sa révision de 1992, [2] dispose ainsi que *le français est la langue de la République*. En Inde, fédération d'états, les frontières de certains états ont bougé depuis l'indépendance (1947) pour s'ajuster à l'évolution des fontières linguistiques, et au moins un état nouveau a été créé, celui de Goa, sur la base du conkani.

1. Office Public de la Langue Bretonne, `www.fr.opab-oplb.org/`
2. `http://www.conseil-constitutionnel.fr/conseil-constitutionnel/francais/la-constitution/les-revisions-constitutionnelles/loi-constitutionnelle-n-92-554-du-25-juin-1992.138025.html`

La politique d'uniformisation linguistique en France a mené non seulement à imposer que tous utilisent le français, mais aussi, jusqu'à la dernière guerre au moins, à réprimer l'usage des langues dites "régionales", comme le breton, l'alsacien, l'occitan, le catalan ou le basque. Pourtant les langues sont intrinsèquement et d'abord liées aux divers patrimoines culturels, et les faire mourir reviendrait à détruire une partie importante de l'identité culturelle de nombreux citoyens.

De fait, Claude Hagège commence sa présentation de la linguistique générale [3] en disant que *le langage est une faculté définitoire de l'être humain*, et que *les langues sont la manifestation historique et sociale de cette faculté*. Les langues (et leurs systèmes d'écriture, avec les calligraphies associées) semblent donc être d'abord inséparables des *cultures* qui les ont engendrées et qu'elles ont nourries et nourrissent, et ensuite (et parfois) seulement être liées à une identité nationale.

Il est donc tout à fait normal que les citoyens et plus généralement les habitants d'un pays comme la France, ayant une langue officielle unique, aient un sentiment aigu d'appartenance à telle ou telle culture dont ils sont issus et à laquelle ils participent, et souhaitent approfondir leur connaissance de leur *langue d'héritage*, ou au minimum comprendre comment il se fait que leur façon d'utiliser leur *langue d'usage* soit influencée par cette langue d'héritage, sans même qu'ils la pratiquent, et qu'ils se reconnaissent immédiatement entre participants de la même *culture d'héritage*, alors même qu'ils se parlent dans la langue d'usage. C'est ce qui se passe pour les Bretons en France.

Le projet *Akenou-Breizh* que nous présentons ici est porté par des informaticiens et des informaticiens-linguistes qui éprouvent ce besoin de mieux comprendre la part de leur identité culturelle liée au breton, langue que la plupart ne parlent pas et n'ont que peu ou pas entendue. Quelques-uns voudraient devenir des auto-apprenants du breton et utiliser pour cela des outils et des ressources intégrés au Web contributif (2.0), et au Web sémantique (3.0), en contexte multilingue. Tous voudraient aussi concilier des aspects touchant à leur recherche, et des aspects génériques concernant le "soutien aux langues peu dotées". C'est dire que le projet, s'il commence concrètement sur le breton-français, se veut ouvert à des participants qui s'intéresseraient à d'autres couples *langue d'héritage – langue d'usage*, comme amazigh-arabe, arabe-français, comorien-français, français-anglais (au Canada ou en Acadie), irlandais-anglais, basque-français, etc.

Plus concrètement, le projet *Akenou-Breizh* vise (1) à mettre en place une plate-forme permettant d'étudier les influences d'une *langue d'héritage*, comme le breton, sur une *langue d'usage*, comme le français, et (2) à proposer à l'intention de tous les intéressés des outils s'intégrant au "Web sémantique et multilingue", et proposant des accès proactifs aux connaissances sur le breton, ainsi qu'une visualisation directe des correspondances sous-phrastiques dans des présentations bilingues alignées (voir figure 1a). Nous nous proposons non seulement d'utiliser les nombreuses ressources disponibles librement, en particulier celles de l'OPLB et du projet Apertium, mais aussi d'en créer de nouvelles, comme des corpus bilingues alignés de bonne qualité, en utilisant le "Web collaboratif", et de construire sur le site dédié `lingwarium.org` [4] des modules linguistiques améliorant ou étendant ceux qui existent, par exemple un analyseur-générateur morphologique.

Nous décrivons aussi une expérience réalisée à partir d'un lexique réduit pour le breton, qui montre comment on peut enrichir un dictionnaire classique, en le reliant à un treillis de thèmes et à un système de gestion de contexte (ici CAMELIS), de façon à ce qu'on puisse l'interroger selon des facettes sémantiques variées, et comparer différentes ressources lexicales après les avoir enrichies de cette façon.

Nous commencerons bien sûr par présenter brièvement l'état de l'art, c'est à dire les ressources, les outils et les plates-formes à accès ouvert concernant spécifiquement le breton-français, ou génériques et utilisables dans ce contexte. Dans la section suivante, nous présentons la méthodologie prévue pour le projet *Akenou-Breizh* : (1) à quelles ressources s'intéresser, comment les utiliser, les étendre, les intégrer, et (2) quelles recherches mener durant le projet en utilisant ces ressources. Dans la dernière section, nous présentons en détail l'expérience d'enrichissement d'un dictionnaire du breton et les possibilités qu'elle ouvre pour faire de l'accès lexical selon des facettes sémantiques variées.

3. http ://claude.hagege.free.fr/

4. Le site `apertium.org` offre des outils de développement de systèmes de TA "à transfert de surface" pour des couples de langues à structures de surface très voisines, comme espagnol-catalan ou espagnol-galicien. Ces outils ne permettent cependant pas d'obtenir de bonnes traductions pour des couples éloignés comme russe-français ou français-breton (pour lesquels il faut disposer d'outils permettant de passer par des structures "abstraites"), ni d'augmenter la qualité par spécialisation heuristique à des sous-langages. Le site `http://www.lingwarium.org` créé par Vincent Berment est similaire, mais offre tous les langages spécialisés (ATEF, ROBRA, TRACOMPL, TRANSF/EXPANS, SYGMOR) du système Ariane-G5 créé par Bernard Vauquois et son équipe entre 1971 et 1985, et tous les modules et systèmes de TA réalisés avec eux entre 1972 et 2015, en source ouvert (ru-fr, fr/pr-en, en-my/th, UNL-fr, maquettes zh-fr/en/de/jp/ru et BEX-FEX). Il est prévu de lui ajouter d'autres outils obtenus par réingénierie, comme le langage des SYSTÈMES-Q d'Alain Colmerauer avec lequel le système TAUM-MÉTÉO fut réalisé en 1977 (après 1985, il fut réécrit en GRAMR par J. Chandioux).

2 État de l'art

2.1 Approches informatiques liées au projet

Les progrès de ces dernières années dans les technologies du Web permettent de proposer à tous les internautes des interfaces d'accès à des informations et à des outils répartis sur de nombreux serveurs, et cela à travers les navigateurs installés sur microordinateurs, tablettes et téléphones mobiles.

Accès et visualisation d'informations, utilisation proactive d'outils. La technique des *passerelles* (gateways) permet, d'enrichir l'interface usuelle de lecture d'un texte à travers le Web de diverses façons. L'une des plus intéressantes consiste à faire surgir une *palette* (relative à un mot ou à une phrase) quand le curseur passe au-dessus d'un mot ou d'une phrase. Une *palette de mot* peut contenir des informations morphosyntaxiques (par exemple, pour "irais" : lemme = "aller", cat = "VRB", mode = "COND", temps = "PRES", personne = "1 2", nombre = "SIN"...), ainsi éventuellement que des indications sur son ou ses sens (par exemple, pour "charme" comme NOM : sens = "#arbre #qualité-de-personne #propriété-de-quark"). Il est également possible de présenter les informations trouvées dans plusieurs dictionnaires et bases terminologiques multilingues sous la forme d'un article de dictionnaire construit à partir de ces informations. C'est le principe de l'outil ALEXANDRIA de Dominique Dutoit [5].

Une *palette de phrase* peut contenir une ou plusieurs traductions, une ou plusieurs analyses grammaticales, ou une traduction, avec ou sans une visualisation de l'alignement entre l'original et la traduction. La figure 1a montre un *amphigramme* (alignement hiérarchique entre une phrase source et sa traduction), et la figure 1b illustre la visualisation dynamique produite en SVG par la technique de Christophe Chenon (Chenon, 2005).

Un point important ici est qu'on peut précalculer toutes les informations potentiellement intéressantes pour les utilisateurs à l'avance, et donc ne jamais les faire attendre quand ils demandent une information, une visualisation ou un traitement. C'est une condition nécessaire pour pouvoir proposer des aides *proactives* à la lecture, à l'apprentissage, à l'annotation, ou à des études spécifiques. La *proactivité* est le fait que les fonctionnalités en question peuvent être activées sans que l'utilisateur n'ait à faire aucune action (autre que de régler les paramètres de proactivité bien sûr).

Avec les outils modernes de suivi oculaire par les webcams intégrées à de nombreux dispositifs, on pourrait même produire (dans la page en cours, dans une *palette de paragraphe* ou dans un nouvel onglet) une présentation proactive "à la Vocable" [6] du paragraphe sur lequel l'utilisateur porte son regard : une zone de *minidictionnaire local* serait alors remplie par l'information lexicale liée au paragraphe en question, comme si on avait passé le curseur dessus.

Construction contributive de ressources et d'outils. Nous sommes depuis presque 10 ans dans l'ère du *Web contributif* [7], qui a donné lieu à de nombreux projets de création contributive de ressources linguistiques. En ce qui concerne les **connaissances lexicales**, on peut mentionner les projets PAPILLON, ITOLDU, WIKTIONARY et JEUXDEMOTS.

Lancée en 2001 par Mathieu Mangeot et Gilles Sérasset, la base lexicale multilingue en ligne PAPILLON-CDM regroupait en 2003 une vingtaine de dictionnaires électroniques bilingues ou multicible concernant 9 langues, et plus de 2M d'entrées. Son "socle logiciel" `Jibiki` permet de définir des bases lexicales de structures très variées, la définition d'une telle base consistant en la description de sa *macrostructure*, et de la *microstructure* de chacun de ses *volumes*.

Destiné à des élèves-ingénieurs devant apprendre le vocabulaire technique anglais de leurs spécialités, le service Web ITOLDU (Bellynck *et al.*, 2005) a permis en 2003-04 de collecter 17.000 entrées en anglais→français, correctes à plus de 90%, avec 250 étudiants répartis en 15 classes, et 6 enseignants ; 30% de la note d'anglais était donnée par le système.

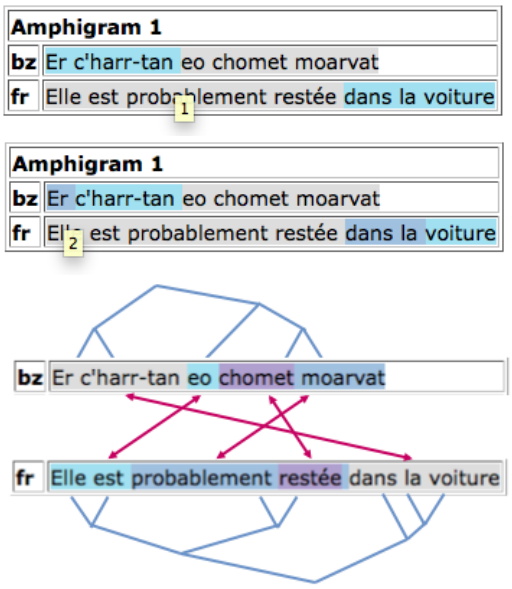
Le projet WIKTIONARY a été lancé en 2002, et contient un dictionnaire d'environ 33.600 mots pour le breton (au 12/4/2015). Voir `http://www.wiktionary.org/` et `http://en.wikipedia.org/wiki/Wiktionary` et `https://br.wiktionary.org/wiki/Wikeriadur:Degemer`.

Il est assez difficile de susciter de nombreuses contributions lexicales de nombreuses personnes. Il y a le plus souvent un petit nombre de "mordus", mais les autres ne contribuent presque rien. Utiliser la myriadisation (Amazon-Turk) ne va pas

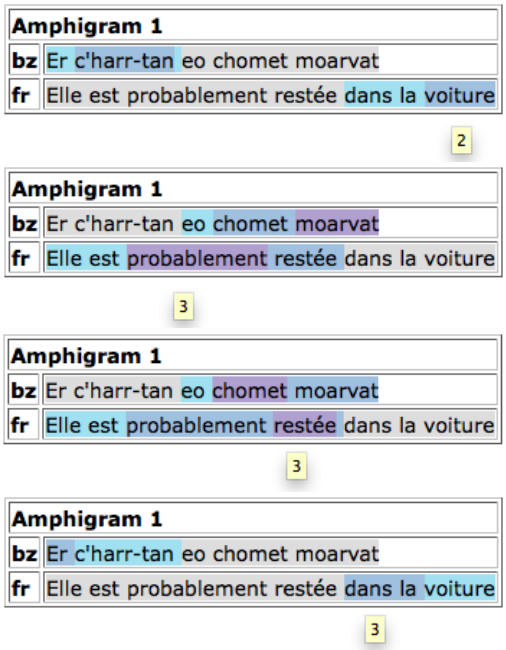
5. Une bonne présentation se trouve à l'url `http://www.tv5monde.com/TV5Site/alexandria/entretien.php` : *Comment peut-on définir le logiciel Alexandria ?*
Alexandria appartient à plusieurs familles de logiciel. Une première famille est celle des logiciels d'aide contextuelle ; en effet, Alexandria fournit un service d'aide à la compréhension (définitions, traductions...) quand un lecteur le sollicite. D'autre part, Alexandria appartient à la famille des agents "intelligents" : ce service est disponible en chaque lieu où l'agent a été installé. Aujourd'hui, les lieux où l'on trouve les réplications les plus courantes de cet agent sont les pages html du Web. C'est le cas pour TV5MONDE.org. Mais Alexandria peut prendre aussi d'autres formes techniques.

6. Voir par exemple `http://www.vocable.fr/pdfreader/magazines/vocD632o.pdf`

7. aussi appelé de façon cryptique *Web 2.0*.



(a) Exemple d'amphigramme breton-français. En fait, l'amphigramme est le bi-arbre n-aire obtenu par *tassement* minimal des deux arbres binaires obtenus à partir des scores des *coupures*, voir (Chenon, 2005).



(b) Visualisation dynamique, exemple en français-breton

FIGURE 1: Visualisation de correspondances sous-phrastiques hiérarchiques (amphigrammes).
On voit deux inversions d'ordre, et aussi une correspondance entre un mot (*er*) et deux mots (*dans la*). Les six rectangles montrent comment la visualisation évolue au fur et à mesure qu'on déplace le curseur, dont la position est indiquée par le petit rectangle jaune numéroté. La couleur la plus foncée correspond au groupe le plus petit contenant le curseur.

non plus : même si on paye très peu, ce serait très cher car on vise plus d'un million de mots (on en a plus dans Wiktionary pour l'anglais et le français), et surtout la qualité des informations serait bien trop basse en moyenne, comme plusieurs expériences l'ont prouvé. [8]

Une idée très intéressante est de *motiver les contributeurs potentiels par un aspect ludique*. C'est ce que fait le système JEUXDEMOTS (voir `http://www.jeuxdemots.org`). Créé par Mathieu Lafourcade et Gilles Sérasset en 2006, ce système a ensuite été développé et considérablement étendu par Mathieu Lafourcade. L'objet de ce projet est la construction de ressources lexicales, et plus précisément d'un réseau lexical du français, pouvant servir à diverses applications du TALN. Ces données sont le produit de l'activité des joueurs de JeuxDeMots. Le réseau lexical du français de JDM contient plusieurs centaines de milliers de mots et de relations, telles que la synonymie ou quasi-synonymie, l'antonymie, l'intensification, et plusieurs fonctions lexicosémantiques (FLS) de Mel'tchuk. Il existe des versions de JeuxDeMots pour d'autres langues que le français ; le projet *Akenou-Breizh* se propose d'en créer une pour le breton.

En ce qui concerne les **corpus**, des outils comme IMAG/SECTRA (Boitet *et al.*, 2010) ont déjà permis de créer de bons *corpus parallèles* par post-édition contributive de résultats de TA. Bien que ces résultats soient souvent très mauvais, s'ils sont jugés par des traducteurs ou des linguistes, ils ont une très grande qualité d'usage, si on les utilise pour faire de la post-édition. Ainsi, deux stagiaires Chinois ont pu, durant un stage d'été en 2013, post-éditer environ 500 pages (10.000 segments) de supports de cours d'informatique, à raison d'environ 10 mn/page. Le résultat n'est pas de qualité professionnelle, mais il est comparable à ce que produirait un traducteur junior en 1 h/page environ, la connaissance du domaine et de la terminologie compensant la différence de niveau en langue source. Il semble que même une traduction "pidgin" [9] puisse permettre de post-éditer une page en 25 à 30 mn dans ces conditions.

8. Voir par exemple `http://www.liberation.fr/societe/2015/05/07/miracles-et-mirages-du-crowdsourcing_1297262` et les autres références de Karën Fort & al. sur le sujet.

9. comme celle produite pour le lao-français sur le site `laosoftware.com` : on donne la ou les traductions possibles de chaque mot, dans l'ordre du texte source, avec éventuellement des annotations utiles pour quelqu'un ayant quelques rudiments de la langue source, comme par exemple l'existence

Possibilité d'amélioration contributive des ressources pendant une activité différente. La post-édition de traductions brutes est faite assez volontiers par des internautes accédant des pages Web "prétraduites" dans leur langue, du moment que l'interface est "sans couture" (sans changement de contexte, i.e. sans nouvel onglet ni nouvelle fenêtre).

Il faut également mentionner la possibilité d'améliorer des annotations de tous types de façon contributive, en parallèle à une activité voulue, comme la lecture. Il peut s'agir de sens des mots, d'arbres syntaxiques, d'alignements, ou de graphes sémantiques. La condition essentielle pour y arriver est que l'utilisateur dispose d'une interface intuitive, proactive, à manipulation directe [10].

2.2 Outils et ressources propres au breton

Plusieurs sites, relevant de diverses initiatives (institutionnelles, associatives ou individuelles), regroupent des liens pour le breton, nous en citons quelques-uns : `http://www.fr.opab-oplb.org/35-outils-linguistiques.htm`, par l'office public de la langue bretonne (OPLB), et `http://www.lexilogos.com`. Nous indiquons ci-dessous des outils et ressources liés à notre projet sur le breton.

2.2.1 Logiciels pour le breton

Traduction (APERTIUM et OPLB). Apertium propose une plateforme libre de droits pour développer la traduction automatique entre des paires de langues. Pour le breton, le site de l'OPLB propose un outil en ligne de traduction automatique dans le sens breton→français. L'OPLB développe ce traducteur, en collaboration avec APERTIUM. L'outil `glosbe` repris plus loin propose certaines traductions, dans les deux sens entre le français et le breton.

Détection de langue. l'identification de la langue est parfois une première étape ; une méthode pour les langues celtiques est proposée dans (Minocha & Tyers, 2014). Il y a peu de détecteurs accessibles, comme `openxerox.com` ou `G2LI`, qui gèrent le breton.

Correcteur grammatical. `https://www.languagetool.org/` propose des correcteurs grammaticaux, en open-source, avec un test en ligne, pour plusieurs en langues dont le breton. L'association `http://www.drouizig.org` propose un correcteur pour le breton, `An Drouizig Difazier`, pouvant être intégré à `Office`.

Il faut noter que les langues celtiques présentent certaines particularités, comme le phénomène des mutations (variations de la consonne initiale d'un mot, voir plus bas pour des détails sur le breton), dont une approche possible est décrite par (Poibeau, 2014). La liste d'outils en ligne ci-dessus n'est pas exhaustive, mais nous constatons certains manques, par exemple concernant l'analyse morphologique.

2.2.2 Ressources pour le breton

Dictionnaires en ligne

`Meurgorf` est un dictionnaire historique du breton en ligne (interrogeable) proposé par l'OPBL. Il contient actuellement 55.747 entrées [11], et continue à s'enrichir (`Meurgorf` est un projet).

APERTIUM et l'OPLB proposent plusieurs dictionnaires concernant le breton disponibles dans APERTIUM (Tyers *et al.*, 2009; Forcada *et al.*, 2011) (dans un format XML spécifique). Il s'agit d'un dictionnaire morphologique par langue (pour le breton, pour le français), et d'un dictionnaire bilingue portant des indications grammaticales pour les mots des deux langues.

Francis Favereau est l'auteur de plusieurs dictionnaires, avec des versions en ligne `http://www.arkaevraz.net/dicobzh/index.php` (volumétrie : fr : 37.401 ; bz : 33.440) permettant d'ajuster la recherche (par exemple, avec mutation : "vugale" trouve "bugale").

Un autre lien est : `http://www.agencebretagnepresse.com/cgi-bin/dico.cgi`.

Tomaz Jacquet est l'auteur du dictionnaire `Freelang`, dans les deux sens entre le français et le breton, voir `http://www.freelang.com/enligne/breton.php?lg=fr` (volumétrie : 37.800 entrées).

`Brezhoneg 21` `http://www.brezhoneg21.com` est une ressource scolaire, en sciences et techniques.

de cas en russe, pour du russe-français.

10. comme par exemple le logiciel *Annot*^{ED} de Johan Ségura (Ségura, 2012).

11. `http://meurgorf.opab-oplb.org/page/index/pr__sentation_du_projet`, consulté le 8/4/2015.

Geriadur, disponible à http://www.geriadur.com/, traduit du français vers le breton (volumétrie : traduction de 22.302 mots français).

Logos gère une ressource libre et contributive, avec des dictionnaires monolingues ou multilingues, interrogeables (à http://www.logosdictionary.org/index.php). Il comprend une version monolingue pour le breton.

http://br.wiktionary.org/ : le projet WIKTIONARY de dictionnaires descriptifs et libres contient un sous-projet pour le breton.

https://fr.glosbe.com/br/fr : ce site se présente comme un dictionnaire multilingue. Il propose des traductions (par exemple entre le français et le breton), et utilise des *mémoires de traductions*. Par exemple, en réponse à la question "pajennoù", le site glosbe répondra que le mot est inconnu mais que l'expression similaire "pajennoù melen" est connue (avec l'équivalent : "pages jaunes").

Autres ressources

Le site ARBRES (http://arbres.iker.cnrs.fr/index.php/Arbres:Le_site_de_grammaire_du_breton) est un site d'informations sur la grammaire du breton, et un centre de ressources pour la recherche en syntaxe formelle sur la langue bretonne (Jouitteau, 2005).

Mentionnons aussi l'atlas linguistique ALBB (http://sbahuaud.free.fr/ALBB/), la base de données toponymique KerOfis de l'OPLB (http://www.fr.opab-oplb.org/40-kerofis.htm), et la base de données TermOfis du centre de terminologie de l'OPLB, avec 62.794 termes (http://www.fr.opab-oplb.org/36-termofis.htm.

Enfin, on peut consulter l'inventaire fait par par l'ELDA, en 2014 [12].

3 Méthodologie

La méthodologie du projet *Akenou-Breizh* s'articule autour de deux pôles : *ressources*, et *recherches*. Les ressources en question sont bien sûr celles qu'on estime nécessaires pour mener les recherches qui, *in fine*, motivent le projet. Cependant, on ne peut pas disposer des ressources très détaillées ou très volumineuses requises par certaines recherches, même si ce sont les plus intéressantes, avant de disposer de ressources moins détaillées ou moins volumineuses. C'est pourquoi la méthode suivie par le projet consistera à élaborer des ressources, des annotations, des outils d'accès ou des traitements au fur et à mesure des possibilités, et à lancer les recherches souhaitées quand les ressources et outils minimaux nécessaires seront disponibles.

3.1 Construction ou collecte de ressources, facilitation d'accès, enrichissement

Corpus parallèles avec visualisation des alignements sous-phrastiques. La pemière tâche concrète du projet sera la construction d'une base de données (BDcorp) contenant des *corpus parallèles* (phrase à phrase en regard) français-breton, munis d'accès interactifs à des dictionnaires en ligne (ce qu'on sait déjà bien faire, grâce à des outils comme ALEXANDRIA ou IMAG/SECTRA) et montrant par des couleurs les correspondances sous-phrastiques (voir figure 1 ci-dessus), l'infrastructure de la BDcorp devant permettre d'ajouter des annotations variées (comme des arbres linguistiques, des annotations discursives, etc.), et de mener des études visant à construire et exploiter ces annotations.

Pour construire des corpus parallèles (alignés au niveau des phrases), et cela de façon essentiellement *contributive (et bénévole)*, on commencera par créer les mémoires de traductions associées, et à améliorer les traductions, si nécessaire, en utilisant une passerelle IMAG/SECTRA d'accès multilingue interactif, qui permet à des contributeurs organisés ou occasionnels de "post-éditer" les traductions, segment par segment (Boitet *et al.*, 2010) [13]. Il est aussi possible de créer de nouvelles traductions en utilisant cette même interface : on peut partir de traductions "pidgin" (cf. supra), reposant sur un analyseur morphologique, un dictionnaire bilingue, et un générateur morphologique. Ces trois composants existent déjà pour le breton, en source ouvert (dans le projet APERTIUM), ce qui fournit un point de départ immédiatement utilisable.

12. Rapport disponible à http://www.culturecommunication.gouv.fr/content/download/106817/1248227/version/1/file/Rapport_dglflf_05112014.pdf

13. http://service.aximag.fr/xwiki/bin/view/imag/home

Il y a ici une possibilité intéressante, qui n'est pas de la recherche, mais une aide potentielle importante à l'auto-apprentissage ou à la simple découverte du breton. Elle consiste à produire dynamiquement une vue bilingue parallèle avec visualisation des correspondances sous-phrastiques, ou, pour les plus avancés, une vue monolingue annotée *à la Vocable*, c'est-à-dire avec affichage proactif d'un minidictionnaire associé à la page, au paragraphe ou au segment (phrase ou titre) en cours de lecture. De même que pour les traductions, il est possible de corriger interactivement les alignements, par manipulation directe depuis le contexte de lecture (Ségura, 2012).

Quand on disposera de suffisamment de bons bisegments pour un sous-langage donné [14] (par exemple, des sites Web ou des romans sur la Bretagne, ou des sections de journaux), on pourra développer des systèmes "empiriques" de TA (statistiques ou fondés sur les exemples) en utilisant des outils comme Moses (en version "factorielle" à cause de la richesse des morphologies flexionnelles du breton et du français, en intégrant les mutations dans la morphologie flexionnelle).

Intégration du breton dans une base lexicale multilingue organisée par acceptions interlingues. Une seconde tâche essentielle est de construire une base lexicale plus "sémantique" que ce qui existe, bien sûr en commençant par réutiliser l'existant, et si possible en le faisant par accès aux ressources en ligne, de façon à bénéficier immédiatement des apports dont elles-mêmes bénéficient constamment.

La première étape serait sans doute d'importer les dictionnaires libres de droits et existant en format XML dans la base lexicale multilingue Papillon-CDM créée et gérée par Mathieu Mangeot (Mangeot, 2001) [15]. La seconde consisterait à enrichir ces données au niveau sémantique, en les reliant à un treillis de domaines et à un ensemble de dénotations de sens, comme les lexèmes interlingues (dits UW) d'UNL [16] (voir `http://www.undl.org`).

Amélioration et extension d'outils pour l'analyse et la génération morphologique. Pour pouvoir accéder aux informations lexicales à partir des mots des textes, il faut en faire l'analyse morphologique. Cela peut se faire par un analyseur, activé pour chaque forme, ou par consultation d'une liste de formes accompagnées de leurs attributs morphosyntaxiques, produite par génération morphologique.

Le projet *Akenou-Breizh* cherchera à étendre les outils existants, tant en termes de couverture que de puissance : il s'agit de couvrir tous les mots simples des dictionnaires existants, de traiter la morphologie dérivationnelle et la morphologie compositionnelle (pour les mots composés), et enfin de construire une "grammaire du mot inconnu" (un "devineur") pour le breton, comme celle réalisée en ATEF [17] par J. Ph. Guilbaud pour le français (Guilbaud & Boitet, 1997).

Désambiguïsation lexicale et préparation à l'extraction de contenu. On utilisera les ressources lexico-sémantiques produites pour transposer au breton les techniques de désambiguïsation lexicale (WSD) qui fonctionnent déjà pour les langues "bien dotées" (le domaine a bien progressé ces dernières années, en particulier grâce à l'organisation des campagnes CLEF). Cela permettra d'intégrer le breton à un extracteur multilingue de contenu construit sur le modèle proposé par le projet ANR OMNIA (Falaise *et al.*, 2010). [18].

3.2 Recherches utilisant ces ressources

Les recherches envisagées par le projet *Akenou-Breizh* devraient concourir à l'étude de l'influence des langues d'héritage sur les langues d'usage. On peut distinger celles qui portent sur les phénomènes linguistiques, à plusieurs niveaux d'interprétation, et à plusieurs degrés de complexité, de celles qui toucheront aux aspects plus pragmatiques et culturels des rapports entre breton et français.

14. Pour un sous-langage assez restreint, partir de 2.000 bisegments de 20 mots, soit environ 100 pages standard, est suffisant si on fait l'apprentissage sur un corpus annoté (par les valeurs des "facteurs", i.e. des attributs morphosyntaxiques).

15. Voir `http://www.papillon.org`. L'import se fait en quelques minutes, une fois qu'on a précisé par des chemins `Xpath` où est l'information correspondant à chaque balise CDM (*Common Dictionary Markup*).

16. UNL = Universal Networking Language. C'est à la fois un projet initié par l'UNU en 1996 et un formalisme de graphes "anglo-sémantiques" utilisable tant pour la traduction que pour le résumé et l'extraction de contenu.

17. ATEF = Analyse de Textes en États Finis. C'est un langage créé par J. Chauché en 1975 (voir Microfiche ACL de 1975) pour écrire des analyseurs morphologiques. C'est le tout premier langage spécialisé fondé sur le modèle des transducteurs finis. Il contient des extensions permettant la programmation heuristique, la modification dynamique de la chaîne en entrée, l'analyse de mots composés, et un vrai traitement des mots inconnu.

18. Ce modèle est en plusieurs étapes. (1) On transforme une phrase ou un texte (ou la partie textuelle d'une requête) en un graphe donnant les segmentations et les analyses morphologiques possibles. (2) On enrichit ce graphe en attachant à chaque lemme tous les UW (lexèmes interlingues UNL) lui correspondant dans la base lexicale. (3) On utilise un module de désambiguïsation lexicale qui attache des scores aux UW. (4) On construit ou on met à jour (automatiquement) un alignement entre la "préontologie" formée par l'ensemble des UW et l'ontologie (ou la base de connaissances) vers laquelle on veut extraire le contenu "pertinent". (5) À l'aide d'une connaissance minimale de la langue du texte (grammaires locales), on construit et on score les groupes élémentaires (*chunks*) formé d'UW alignés avec des concepts ou des attributs de l'ontologie, et on produit un *descripteur* (une liste attributs-valeurs, ou une liste de petits graphes UNL). (6) On transforme ce descripteur en une *description* dans le langage de l'ontologie. Si c'est une donnée, on la range dans la T-box (conteneur des *termes*). Si c'est une requête, on lance l'agorithme de production d'une réponse.

3.2.1 Recherches à plusieurs niveaux et à plusieurs degrés de complexité

Il s'agit de

- recherches simples basées sur les textes parallèles ;
- recherches nécessitant aussi des dictionnaires (donc sans doute de la lemmatisation) :
  . extraction de mots ou termes hors dictionnaire,
  . extraction de couples de mots ou de termes "candidats" à l'équivalence (en traduction),
  . étude des registres d'expression (par exemple, formes polies d'un côté, directes de l'autre).
- recherches nécessitant aussi des analyses sous forme d'arbres syntaxiques :
  . oppositions actif/passif,
  . expressions de la modalité et de la politesse (tournures, modes...),
  . étude contrastive des "mots composés" ou "tournures".

3.2.2 Recherches autour de la pragmatique du breton (approche statistique et linguistique)

Au travers de présupposés et sous-entendus portés par des locutions choisies, on vise

- à étudier les associations de termes et leurs variations dans le temps et l'espace,
- à comparer avec le français et à mettre en évidence des différences et influences entre les deux langues.

4 Un cas d'étude

Nous décrivons une expérience qui met en œuvre un petit lexique et un système d'information permettant de le manipuler.

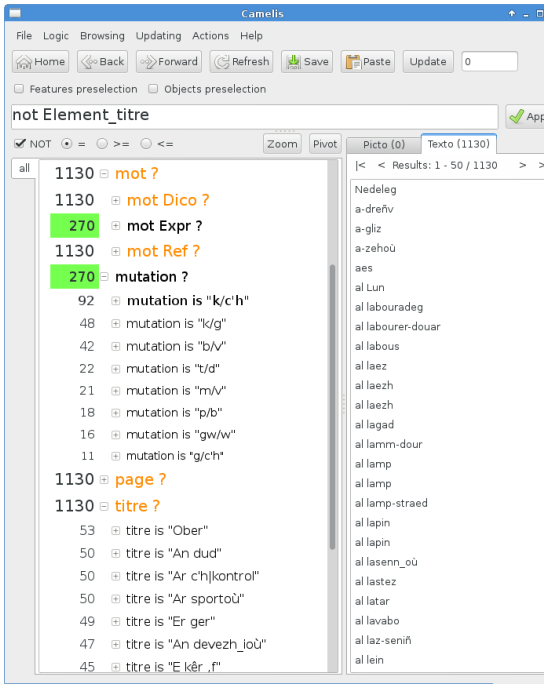
4.1 Ressources et outils utilisés

Le lexique choisi suit l'album (Kergoat *et al.*, 2007) qui fournit le vocabulaire fondamental du breton (1000 mots), conçu au départ pour les enfants (pour plus de dix langues). Le lexique a été saisi par thèmes. Il est organisé en conservant le rattachement des mots et expressions à des thèmes correspondant à des scènes de la vie quotidienne ("à la maison", "à l'école", etc.).

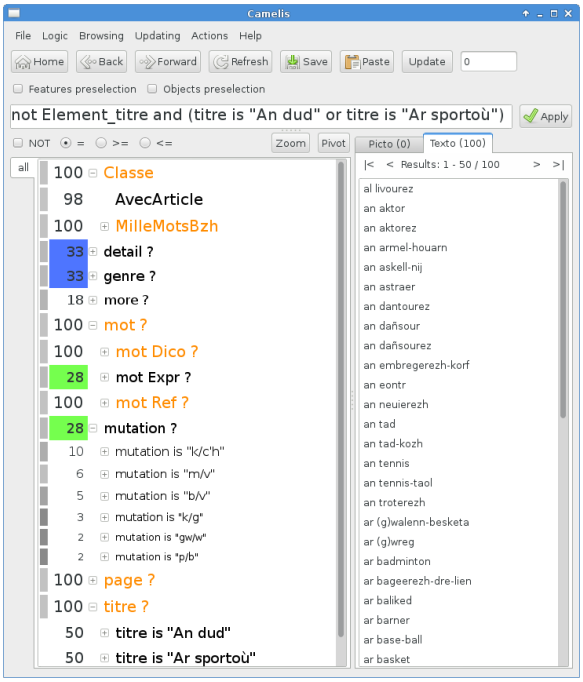
Les noms sont indiqués avec un déterminant, et leur genre (dans l'album aussi). Des synonymes sont proposés dans l'album ; nous les avons aussi indiqués, avec un marquage. Le breton partage avec d'autres langues celtiques un phénomène connu sous le nom de *mutation* : il s'agit d'une modification de la consonne intiale (comme $k > c'h$), régie par plusieurs sortes de déclencheurs [19]. Nous avons choisi d'inclure en plus dans le lexique une indication de mutation (par exemple, l'expression *an daol* pour "une table" sera indiquée par : *an d\|taol*, le lemme du nom étant *taol*, et la mutation avec cet article *an* étant $d > t$. Cette information supplémentaire permettra ensuite d'interroger le *contexte* du lexique selon ce critère (fréquence d'une mutation particulière, son contexte, etc.), éventuellement combiné à d'autres critères.

Le système d'information choisi est un *contexte* CAMELIS, un tel contexte étant défini par un ensemble fini d'objets, avec pour chaque objet un ensemble fini de descriptions (formules logiques). CAMELIS (version 1, accessible à `http://www.irisa.fr/LIS/ferre/camelis/`) est un *système de gestion de contexte logique* permettant plusieurs formes de manipulations flexibles par *facettes* : interrogation/navigation sans connaissance *a priori* (par clics et sélections successives), mais guidé par un index contextuel de propriétés et de manipulations plus expertes (écriture de requêtes, mise à jour). Ce logiciel est basé sur l'analyse de concept logique (LCA) définie dans (Ferré & Ridoux, 2004), qui propose une extension de l'analyse de concept formel (FCA, voir (Ganter & Wille, 1999)). Un *concept logique*, noté c, est un couple formé d'une extension $ext(c)$ (un ensemble d'objets) et d'une intension $int(c)$ (une formule), tel que les éléments de $ext(c)$ sont exactement ceux qui vérifient $int(c)$. Ces concepts forment un treillis auquel correspond l'*arbre de navigation logique* et incrémentale dans la fenêtre gauche du logiciel. Le logiciel CAMELIS est aussi prévu pour gérer des ensembles d'objets de types différents et plusieurs facettes (graduelles) dans l'arbre de navigation.

19. `http://arbres.iker.cnrs.fr/index.php?title=Les_mutations_consonantiques`, consulté le 01/4/2015



(a) Contexte initial, sans les objets pour les titres



(b) Contexte après nouvelle sélection dans l'index

FIGURE 2: Contexte monolingue

4.2 Construction de contextes et usages

Nous illustrons quelques usages du lexique transformé en contexte à explorer : le lexique breton, d'abord pris isolément, ensuite complété avec d'autres ressources. Ce type de scénarios illustre une recherche d'information possible, mais aussi une forme d'évaluation d'une ressource en fonction d'une autre.

Rôles des fenêtres CAMELIS par rapport à un contexte. L'outil CAMELIS, chargé avec un contexte initial, présente trois fenêtres relatives à un contexte courant, qui évolue au fil des sélections dans ces fenêtres. Pour les figures 2a et 2b , le contexte initial ("home", pour "all" dans la fenêtre supérieure) contient deux sortes d'objets, d'une part pour les expressions illustrées dans l'album, et d'autre part pour les titres. Il s'agit au départ d'informations monolingues. Ce contexte "home" sera augmenté par la suite, pour intégrer des informations multilingues. [20]

Fenêtre d'objets : la partie droite présente les objets du contexte courant, par leur label [21]

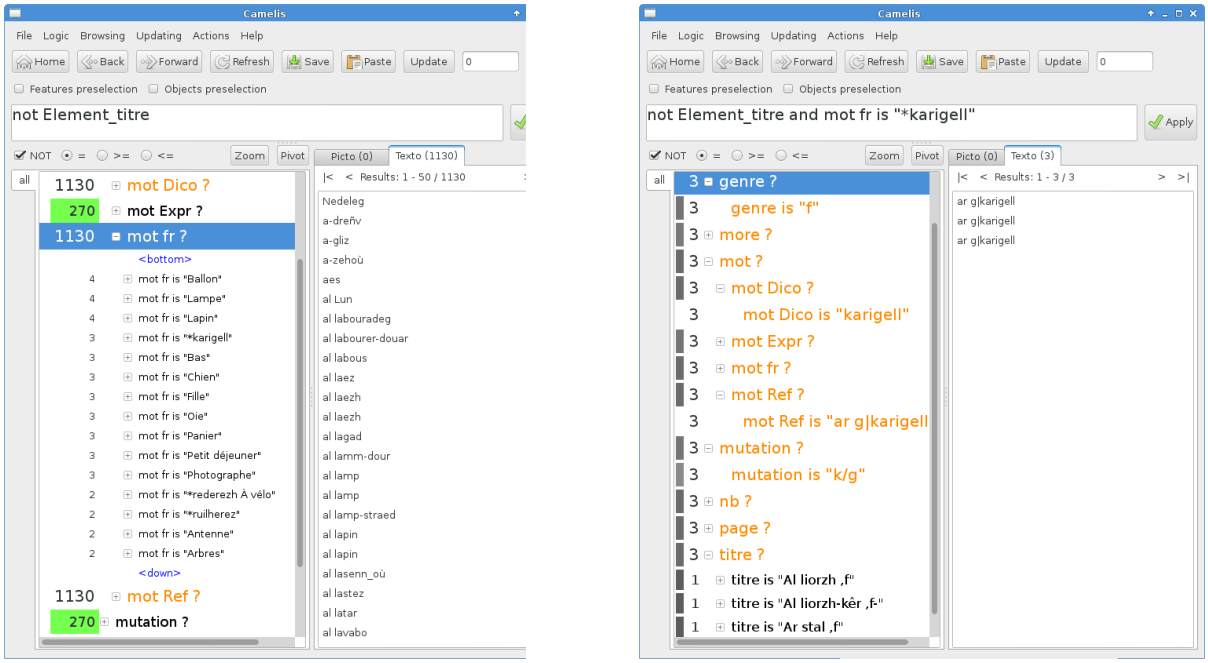
Fenêtre de propriétés : la partie gauche indique les propriétés, organisées en arbres selon les relations entre les propriétés. Il s'agit aussi d'un index cliquable qui permet de passer d'un contexte à un autre.

Fenêtre de requête : la partie du haut contient une requête caractérisant le contexte courant : c'est une propriété satisfaite par tous les objets du contexte courant ; elle n'a pas besoin d'être saisie puisqu'elle est mise à jour automatiquement selon les sélections dans les deux autres fenêtres. L'utilisation ne nécessite pas de connaissance *a priori*, mais il est aussi possible de rédiger directement les requêtes.

Sur le lexique initial

Les thèmes : nous voyons dans l'index de navigation à gauche, la facette `titre` qui indique, *pour chaque objet* mentionné à droite, le titre auquel il est rattaché (dans l'album et dans le contexte). Après sélection à gauche d'un

20. Un tel contexte peut être hétérogène, il peut contenir plus de sortes d'objets et de propriétés, selon les préférences et les usages prévus.
21. ils pourraient être présentés avec une photo, par l'onglet "Picto", selon le type de contexte



(a) Contexte breton et français

(b) Contexte après nouvelle sélection dans l'index

FIGURE 3: Contexte multilingue

`titre` particulier (ou de plusieurs titres, comme dans la figure 2b), les trois fenêtres seront synchonisées pour représenter le nouveau contexte, de façon que les objets à droite soient ceux rattachés à ce `titre` et que la requête en haut représente la propriété choisie, vérifiée par ces objets.

Les mutations : nous voyons aussi plus haut à gauche, la facette `mutation` qui indique, *pour certains objets*, la forme de mutation associée (dans le contexte [22]). Les mutations sont ordonnées ici selon leurs fréquences (on constate que les mutations en "k" dominent). Les couleurs à gauche servent à indiquer des propriétés satisfaites par un même ensemble d'objets : ici, les objets présentant une mutation ont ainsi été annotés pour indiquer leur mot de référence (`mot Ref is "..."`)

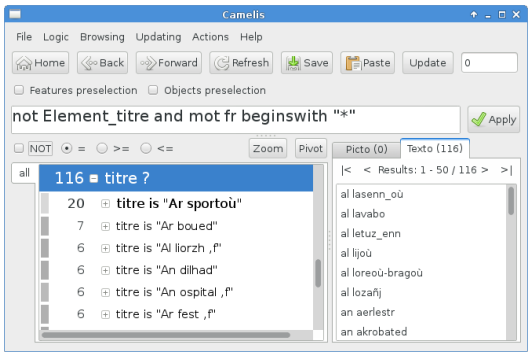
Quelques autres propriétés sont indiquées. Ainsi, la facette `AvecArticle` (sous `Classe` par un "axiome CAMELIS") rend visible le fait que la plupart des expressions contiennent un article, grâce à l'affichage automatique de son cardinal.

Avec intégration de ressources APERTIUM. Le contexte "home" est maintenant augmenté, pour intégrer des informations multilingues provenant d'APERTIUM. Pour chaque langue ajoutée, les mots qui sont inconnus d'APERTIUM sont marqués par * (au début). Cela peut servir de point de départ à diverses évaluations, comme (1) une analyse de couverture du traducteur et des suggestions de complétion, et (2) une caractérisation des cas sans traduction [23]. Certaines incohérences peuvent aussi être mises en évidence.

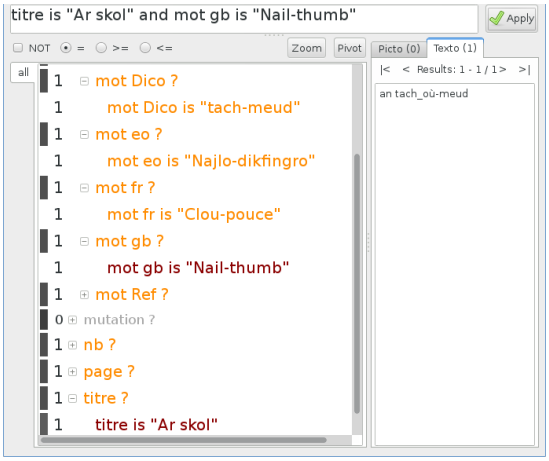
Facettes multilingues. les informations pour le français sont de deux sortes : une propriété *nb* =... indique le nombre d'occurrences dans un fichier dictionnaire pour la paire de langue breton-français ; une propriété `mot fr is ...`, comme dans l'image 3a, indique la traduction fournie par APERTIUM pour le `mot Dico is ...` associé à un objet/expression en breton. Par la suite, la traduction en espéranto avec une facette `mot eo is ...`, puis celle en anglais avec `mot gb is ...` sont aussi ajoutées.

Mise en évidence de manques. L'image 3b illustre un cas de mot breton présent dans plusieurs pages/thèmes de l'album et pourtant absent d'APERTIUM. Une requête plus fine, comme celle de l'image 4a, permet de constater le nombre

22. ajout par rapport à l'album
23. De cette façon, des erreurs dans une première version du source ont déjà pu être repérées.



(a) Contexte breton et français ; mots inconnus



(b) Contexte br, eo, fr, gb ; thème "école"

FIGURE 4: Contexte multilingue et Apertium

de mots de l'album non connus dans APERTIUM, avec les thématiques principalement concernées (ici les sports : "Ar sportoù"). Globalement, un peu plus de 100 mots parmi ceux de l'album n'ont pas de traduction. [24].

Autres défauts. Des problèmes particuliers peuvent apparaître avec des mots composés. En ajoutant les autres langues, comme dans l'image 4b, les manières d'interroger se multiplient, et une mise en rapport de facettes sans passer par le breton permet d'autres repérages, comme entre le français "clou-pouce" et l'anglais "thumb-nail", associés au même objet.

Note. Toutes les fonctionnalités de CAMELIS ne sont pas exploitées ici, comme les actions et la mise à jour.

5 Conclusion

Nous avons présenté dans cet article un nouveau projet, *Akenou-Breizh*, qui concerne directement le TALN appliqué au breton et au couple breton-français. Plus précisément, il vise (1) à mettre en place une plate-forme permettant d'étudier les influences d'une *langue d'héritage*, comme le breton, sur une *langue d'usage*, comme le français, et (2) à mettre à disposition de tous les intéressés des outils s'intégrant au "Web sémantique et multilingue", et proposant des accès proactifs aux connaissances sur le breton ainsi qu'une visualisation directe des correspondances sous-phrastiques dans des présentations bilingues alignées.

Après avoir présenté un état de l'art assez complet des outils et ressources concernant le breton, puis des nouvelles possibilités apportées par le Web en termes d'interfaces non seulement d'accès enrichi, mais aussi de contribution pour l'enrichissement ou la création de ressources et d'outils permettant des applications et des recherches à divers niveaux, de la morphologie à la pragmatique, nous avons décrit la méthodologie prévue dans le projet.

Notre premier but est de valoriser les ressources et les outils pour le breton (écrit), avec des accès proactifs. Nous souhaitons ensuite les évaluer et les enrichir ; une étude a été menée en ce sens pour permettre un accès lexical avec des facettes sémantiques variées. Il est prévu d'associer à ce projet d'autres partenaires spécialistes du breton qui produisent ou soutiennent des développements pour le breton. À terme, d'autres langues pourraient être aussi concernées, comme celles de la famille celtique.

24. Quelques erreurs dans l'album ne sont pas exclues non plus, cette méthode peut aider à les repérer aussi

Remerciements

Nous remercions l'Office public de la langue bretonne et des collègues de Rennes-2, pour les discussions utiles et leur soutien pour ce projet.

Références

BELLYNCK V., BOITET C. & KENWRIGHT J. (2005). *ITOLDU, a Web Service to Pool Technical Lexical Terms in a Learning Environment and Contribute to Multilingual Lexical Databases*. In A. GELBUKH, Ed., *Computational Linguistics and Intelligent Text Processing (Proc. CICLING-2005)*, number 3406 in LNCS, p. 319–327 : Springer.

BOITET C., HUYNH C.-P., NGUYEN H.-T. & BELLYNCK V. (2010). *The iMAG concept : multilingual access gateway to an elected Web site with incremental quality increase through collaborative post-edition of MT pretranslations*. In *Actes de TALN 2010, Montréal, Canada*.

CHENON C. (2005). *Vers une meilleure utilisabilité des mémoires de traductions, fondée sur un alignement sous-phrastique*. PhD thesis, UJF.

FALAISE A., ROUQUET D., SCHWAB D., BOITET C. & BLANCHON H. (2010). *Ontology-driven content extraction using interlingual annotation of texts in the OMNIA project*. In *Proc. CLIA workshop of COLING-2010* : ACL.

FERRÉ S. & RIDOUX O. (2004). *Introduction to logical information systems*. *Inf. Process. Manage.*, **40**(3), 383–419.

FORCADA M. L., GINESTÍ-ROSELL M., NORDFALK J., O'REGAN J., ORTIZ-ROJAS S., PÉREZ-ORTIZ J. A., SÁNCHEZ-MARTÍNEZ F., RAMÍREZ-SÁNCHEZ G. & TYERS F. M. (2011). *Apertium : a free/open-source platform for rule-based machine translation*. *Machine translation*, **25**(2), 127–144.

GANTER B. & WILLE R. (1999). *Formal concept analysis - mathematical foundations*. Springer.

GUILBAUD J.-P. & BOITET C. (1997). *Comment rendre une morphologie robuste du français encore plus robuste en traitant finement les mots inconnus avec les données disponibles*. In *Actes de TALN-97, Grenoble*, p. 12 p. : CLIPS, UJF.

JOUITTEAU M. (2005). *La syntaxe comparée du breton, une enquête sur la périphérie gauche de la phrase bretonne*. PhD thesis, Nantes, France.

KERGOAT L., AMERY H. & CARTWRIGHT S. (2007). *Les 1000 premiers mots en breton*. Skol an Emsav, 8 edition.

MANGEOT M. (2001). *Environnements centralisés et distribués pour lexicographes et lexicologues en contexte multilingue*. PhD thesis, UJF.

MINOCHA A. & TYERS F. (2014). *Subsegmental language detection in Celtic language text*. In *Proceedings of the First Celtic Language Technology Workshop*, p. 76–80, Dublin, Ireland : Association for Computational Linguistics and Dublin City University.

POIBEAU T. (2014). *Processing Mutations in Breton with Finite-State Transducers*. In *Proceedings of the First Celtic Language Technology Workshop*, p. 28–32, Dublin, Ireland : Association for Computational Linguistics and Dublin City University.

SÉGURA J. (2012). *Mémoires partagées d'alignements sous-phrastiques bilingues*. PhD thesis, LIRMM, Université de Montpellier II.

TYERS F. M., DUGAST L. & PARK J. (2009). *Rule-based augmentation of training data in Breton–French statistical machine translation*. *Proceedings of the 13th Annual Conference of the European Association of Machine Translation, EAMT09*, p. 213–218.

Feuille de route pour le développement numérique occitan

Benoît Dazéas

Lo Congrès permanent de la lenga occitana, Château d'Este, BP 326, 64141 Billère cedex

b.dazeas@locongres.org

Résumé

Le Livre blanc de META-NET, un réseau d'experts européens en technologies de la langue, alerte sur le risque « d'extinction numérique » de plusieurs langues européennes et de l'urgence pour elles de se doter rapidement de technologies de support. Cette étude propose également une grille de classification et d'évaluation des ressources et préconise des principes d'action tels que la création massive de données, la mutualisation ou encore le transfert technologique.
Dans ce cadre *Lo Congrès permanent de lenga occitana* a piloté la rédaction d'une feuille de route pour le développement numérique de l'occitan. Le rapport final fait état des ressources existantes et propose une planification de réalisation (2015-2019) des ressources de bases et des outils finaux.
La mise en place de cet ambitieux programme nécessitera la coordination des acteurs de transmission de l'occitan – politiques linguistiques, recherche scientifique et communauté du logiciel libre – ainsi que la mobilisation des différents crédits et fonds européens.

Abtract

Roadmap for the Occitan digital development
The META-NET White Paper Series, issued by the European network of experts in language technology, alerts to the risk of digital extiction for several European languages and the emergency to get equipped with supporting technologies. The study also presents a classification scale and assessment of the ressources and recommends taking actions, like massive data import, resource pooling, or technological tranfert.
According to this, *Lo Congrès permanent de la lenga occitana* steered the drawing up of a roadmap for the Occitan digital development. The final report lists the existing resources and schedules the release of each basic resource and final tools (2015-2019).
The implementation of this ambitious workprogram will require a teamwork of all partners involved in the transmission of Occitan – language policies, scientific research and open-source application community – as well as gathering various credits and European Funds.

Mots-clés : Langues régionales et minoritaires, occitan, feuille de route, ressources langagières, technologies de la langue, politiques linguistiques.
Keywords : Regional and minority languages, Occitan, language ressources, language technologies, language policies.

1 Introduction

1.1 L'occitan, une langue européenne

L'occitan est une langue romane parlée dans trois États de l'Union européenne (France, Espagne, Italie) sur un espace d'environ 150 000 km². Sur les 15 millions d'habitants concernés. Il est difficile d'en dénombrer les locuteurs ; à partir des différentes études[1] conduites ces dernières années, partielles et étalées dans le temps, on situe, selon les sources[2], le nombre de locuteurs entre plusieurs centaines de milliers à plusieurs millions de personnes.

En l'absence de standard imposé officiellement et du fait de la vitalité de certaines variétés dialectales, l'occitan peut être défini comme une langue polynomique, composée de six grandes variétés dialectales[3].

1 A noter l'enquête sociolinguistique conduite par la région Aquitaine en 2008 (http://www.aquitaine.fr/content/download/786/7753/file/Enquete_linguistique(1).pdf) ou encore celle de la région Midi-Pyrénées e ; 2010 (http://www.midipyrenees.fr/IMG/pdf/EnqueteOccitan.pdf).
2 http://www.univ-montp3.fr/uoh/occitan/une_langue/co/module_L_occitan_une%20langue_10.html
3 Auvergnat, gascon, languedocien, limousin, provençal, vivaro-alpin (BEC, 1995).

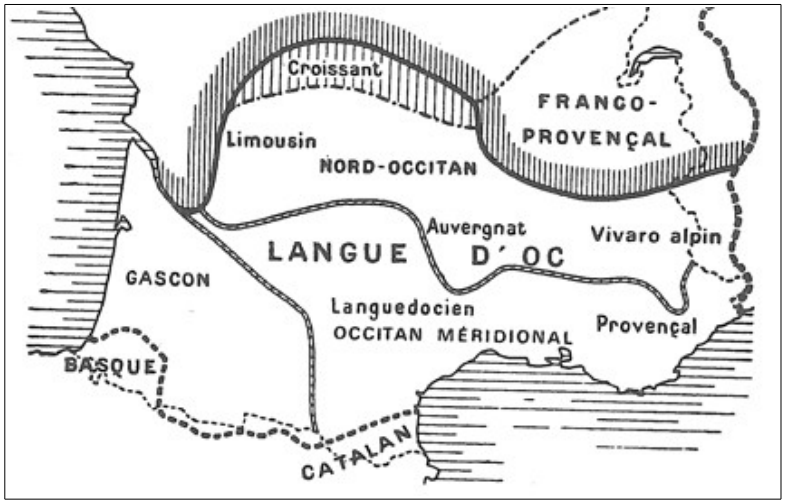


FIGURE 1 – Classification des dialectes occitans selon Pierre Bec

Co-officielle en Val d'Aran, la langue occitane bénéficie, à défaut d'une reconnaissance publique, du soutien des collectivités territoriales en France et fait partie des langues protégées par la loi sur les minorités linguistiques en Italie.

Riche d'une littérature écrite millénaire, l'occitan est aujourd'hui présent dans la presse, sur Internet et à la télévision. Soutenue par une réseau associatif et institutionnel dense, elle est enseignée de la maternelle (enseignement immersif associatif ou bilingue public) jusqu'à l'Université.

1.2 Lo Congrès, une institution collégiale pour réguler l'occitan

Lo Congrès permanent de la lenga occitana est l'organisme interrégional de régulation de l'occitan. Il rassemble les institutions et fédérations historiques du territoire occitanophone et il est soutenu par la Délégation à la langue française et aux langues de France (ministère de la Culture et de la Communication) et les collectivités territoriales. Installé officiellement à l'hôtel de Région Aquitaine à Bordeaux en décembre 2011, il a pour mission de contribuer à la vitalité et au développement de l'occitan – appelé aussi langue d'oc – en travaillant à sa connaissance et à sa codification par la production des outils concernant les différents aspects de la langue (lexicographie, la lexicologie, la terminologie, la néologie, la phonologie, la graphie, la grammaire et la toponymie).

Lo Congrès possède deux organes assesseurs – le Conseil linguistique[4] (dont le président est Patrick Sauzet, linguiste et professeur à l'Université Toulouse 2) et le Conseil des usagers[5] – et agit selon des principes d'action tels que le respect de l'unité et de la diversité de l'occitan, la stabilité, la représentativité des régions linguistiques du territoire d'Oc, la collégialité des décisions et la diffusion de l'information.

Afin de répondre à la demande urgente des usagers, plus spécifiquement ceux du domaine de l'enseignement et de la formation pour adultes, le Congrès a développé une plate-forme numérique – locongres.org – rassemblant différents outils linguistiques de références : un multidictionnaire occitan (*dicod'Òc*), un conjugateur (*vèrb'Òc*), une base terminologique (*tèrm'Òc*), une base toponymique *(top'Òc*), un corpus textuel ainsi qu'un portail d'accès vers les différentes ressources occitanes en ligne.

Avec plus de 180 000 visites en 2014, le portail numérique *locongres.org* est pensé comme un service public en langue occitane : son accès est gratuit, multiplate-formes (*Windows*, *iOS*, *Androïd*, etc.), les formats libres et les licences

4 La Communauté scientifique est représentée au Congrès par le Conseil linguistique. Ce conseil assesseur est déjà constitué et ses membres sont à l'œuvre sur différents travaux. Toutes les régions occitanes y sont représentées. Le Conseil linguistique a un Président et un bureau élu, ainsi que des commissions qui travaillent pour les besoins du Congrès. Liste des membres : http://www.locongres.org/index.php/fr/lo-congres-fr/le-conseil-linguistique/membres

5 Le Conseil des usagers est un conseil assesseur du Congrès ayant pour fonction de représenter la demande sociale. Il rassemble des personnes qualifiées représentatives de la pratique sociale de la langue et qui sont réparties en trois secteurs : les transmetteurs (enseignement, cours pour adultes et formation professionnelle), les utilisateurs (écrivains, éditeurs, médias) et les institutionnels (opérateurs de politiques publiques).

contributives sont privilégiés.

1.3 Les enjeux du développement numérique

Lo Congrès fait partie d'une dynamique qui a permis au numérique occitan de se développer d'une façon générale ces dernières années : contenus encyclopédiques (*Wikipédia*[6]), patrimoine (*Occitanica*[7], *Sondaqui*[8], *trobadours d'Aquitaine*[9]), médias (*Octele*[10]), réseaux sociaux sont autant de secteurs désormais investis. Toutefois la langue occitane pâtit toujours d'un important retard numérique avec pour conséquence, une absence quasi totale dans des outils désormais courants (bureautique, téléphonie mobile, etc.). La prégnance croissante de ces technologies dans la vie quotidienne (travail, déplacements, consommation, éducation, vie sociale) font des technologies du langage un facteur supplémentaire de marginalisation pour une langue déjà minorisée.

Ce phénomène est décrit et analysé dans une étude réalisée par META-NET, un réseau de recherche rassemblant différentes institutions, universités et centres de recherche et dont la mission principale est la mise en place de fondations technologiques solides pour une Europe multilingue. Son Livre blanc[11] fait un état actuel des ressources et technologies du langage pour trente langues européennes dans six domaines (la traduction automatique, la synthèse et la reconnaissance vocale, la correction orthographique, l'analyse sémantique, l'analyse grammaticale et la génération automatique de texte) et propose également une grille commune de classification et d'évaluation des ressources et outils numériques. Les résultats de l'étude sont particulièrement alarmants : les éditeurs soulignent l'écart croissant entre les « grandes » et les « petites » langues, il est indispensable d'équiper toutes les langues (y compris les plus petites et les moins dotées) des technologies de base nécessaires, sans quoi ces langues sont condamnées à « l'extinction numérique ». Pour ce faire, **l'étude préconise la création massive de données, la mutualisation au niveau européen, le transfert technologique entre les langues, l'interopérabilité des ressources, des outils et des services** (REHM, USZKOREIT, 2002 ; SORIA *et al.*, 2013).

Pour ce qui concerne la France, faisant suite à la proposition d'action n°26 du rapport Jacques ATTALI intitulé *La francophonie et la francophilie, moteurs de croissance durable*[12] et remis au Président de la République en août 2014, la Délégation générale à la langue française et aux langues de France (ministère de la Culture et de la Communication) a souhaité lancer un nouveau volet du programme *Technolangue*[13], dont la première édition a permis, entre 2003 et 2005, d'accompagner plusieurs projets d'outillage en traitement automatisé pour la langue française. Ce nouveau programme vise à compléter l'outillage pour la langue française et à développer de nouvelles technologies de traitement pour les langues de France. Dans ce cadre, la DGLFLF, en partenariat avec le CNRS et ELDA, a organisé les 19 et 20 février 2015 un colloque[14] sur le thème du développement des technologies en faveur des langues régionales de France afin de constituer un atelier de réflexion, réunissant une cinquantaine de participants, dont des experts scientifiques, des représentants des collectivités territoriales et des membres d'associations de soutien aux langues régionales.

La DGLFLF a également réalisé en partenariat avec ELDA un inventaire[15] des ressources linguistiques des langues régionales en reprenant le classement établi dans le Livre Blanc de META-NET (SORIA, MARIANI, 2013). Ces travaux comprennent également une étude de la faisabilité de l'application des technologies : l'occitan, qui serait proche du breton en termes de classification, disposerait ainsi des ressources suffisantes pour développer des outils de traduction automatique ou la correction orthographique.

2 Diagnostic et feuille de route pour le développement numérique de la langue occitane

Parallèlement à ces travaux, *Lo Congrès* a initié une recherche-action associant à la fois ses partenaires publics (DGLFLF et collectivités membres du Congrès[16]) et les opérateurs de recherche, de transmission et de diffusion de la langue (Université, opérateurs de missions publiques, formation professionnelle, médias, etc.) autour de la question de la stratégie de développement numérique pour l'occitan, plus spécifiquement pour le domaine du Traitement Automatique des Langues (TAL). L'objectif était, en huit mois de travaux (avril-novembre 2014), de réaliser un

6 http://oc.wikipedia.org
7 http://www.occitanica.eu
8 http://www.sondaqui.com
9 http://www.trobar-aquitaine.org
10 http://www.octele.com
11 http://www.meta-net.eu/whitepapers/press-release-fr
12 Rapport téléchargeable sur le site de la documentation française : http://www.ladocumentationfrancaise.fr/rapports-publics/144000511/
13 http://www.technolangue.net
14 Programme du colloque : http://tlrf2015.sciencesconf.org/
15 Rapport téléchargeable sur le site de la DGLFLF : http://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
16 Ministère de la Culture et de la Communication-DGLFLF, Régions Aquitaine, Midi-Pyrénées, Languedoc-Roussillon, Départements des Pyrénées-Atlantiques et des Hautes-Pyrénées, ville de Toulouse.

diagnostic (inventaire) des ressources et outils linguistiques existants et une feuille de route de développement 2015-2019. Des experts internationaux des TAL pour les langues basque, bretonne, catalane et gallois ont été également associés aux travaux.

L'inventaire des ressources linguistiques[17], réalisé en collaboration avec ELDA, a été publié en juin 2014 (il reste un chantier ouvert). Avec près de 250 références, il permet de faire le diagnostic suivant :

— Il existe très peu d'outils de technologie de la langue en occitan. Les outils existants se concentrent dans la catégorie des correcteurs orthographiques.

— Les ressources recensées sont plus nombreuses, mais peu peuvent être réutilisées :

> — La plupart des ressources lexicales sont constituées de dictionnaires numérisés, mais la plupart sont anciens ou non validés au niveau linguistique. Les autres demandent un gros traitement avant de pouvoir être utilisés en traitement automatique des langues (TAL).
>
> — Les corpus sont nombreux, mais il faut également les trier en fonction de leur qualité linguistique. Par ailleurs, ils ne sont pas, pour la plupart, directement utilisables pour créer des outils (les corpus oraux ne sont pas transcrits, les corpus textuels ne sont pas annotés).

— Les grammaires sont destinées à une utilisation papier davantage qu'à une utilisation informatique.

Il est donc indispensable de créer des ressources linguistiques de base avant de pouvoir développer des outils à partir de ces ressources.

Ressources linguistiques	**Recensées**	**Utilisables en informatique***
Corpus monolingues de textes	27	2
Corpus monolingues de parole	28	0
Corpus parallèles	1	1
Corpus multimédias et multimodaux	24	0
Lexiques	73	8
Bases terminologiques	21	2
Tesauri, Wordnets, ontologies	1	1
Toponymie	3	3
Grammaires, modèles de langage	30	0
Outils de technologie du langage	**Recensés**	
Reconnaissance de l'écriture	0	
Reconnaissance de la parole	0	
Synthèse vocale	0	
Analyse grammaticale	6	
Analyse sémantique	0	
Génération de texte	0	
Traduction automatique	2	
Recherche et extraction d'information	0	
Autres outils	**Recensés**	
Logiciels disponibles en occitan	6	
Outils numériques pour apprendre l'occitan	4	

FIGURE 2 – Diagnostic du développement numérique de la langue occitane (synthèse)

17 http://inventari.locongres.org

La feuille de route pour le développement numérique occitan 2015-2019[18], basée sur le Livre blanc META-NET, a été présentée officiellement le 28 novembre au Congrès à Billère ; cette dernière propose une stratégie harmonisée sur cinq ans (2015-2019) pour le développement des technologies linguistiques de l'occitan, et fournit un cadre cohérent duquel peuvent dériver des actions concrètes pour sa mise en œuvre. Heureusement, l'occitan n'est pas à un point de développement zéro et les initiatives prises à ce jour devraient constituer un point de départ intéressant.

Il est donc essentiel de réutiliser les ressources et les outils existants, ce qui permettra de cibler les efforts à venir. Cette stratégie, basée sur l'optimisation de la coopération entre les différents acteurs, exigera un effort collectif pour veiller à atteindre les objectifs fixés en 2019.

	2015	**2016**	**2017**	**2018**	**2019**
Ressources linguistiques					
Corpus textuels					
Monolingues					
Corpus spécialisés (10-25)			x*		x**
Corpus web (5 millions)			x		
Parallèle (2-5)				x	
Ressources lexicales					
Base lexicale monolingue	x***		x		
Base lexicale bilingue			x		
Grammaires					
Base grammaticale/syntaxique		x			
Outils linguistiques					
Traitement de la parole					
Ressources pour la reconnaissance de la parole					x
Synthèse vocale					x
Détection automatique de la langue					
Détecteur de l'occitan	x				
Détecteur des variantes de l'occitan		x			
Analyse grammaticale					
Correcteurs orthographiques					
Correcteur orthographique polyvalent (toutes les variantes)		x			
Clavier prédictif et autocorrection			x		
Analyseurs					
Lemmatiseur-analyseur morphologique		x			
Analyseur syntaxique					x
Analyse sémantique					
Base de connaissance lexicale				x	
Traduction automatique					
Traducteurs automatiques					
oc --> fr (toutes les variantes)			x		
fr--> oc					x
Transcripteur automatique entre variantes			x		

18 Rapport complet : http://locongres.org/images/docs/feuille_route_numerique_occitan_fr.pdf

Logiciels					
OS + Applications principales			x		

** Corpus monolingue : première version (10 millions de mots), ** Corpus monolingue : deuxième version (25 millions de mots), *** Base lexicale monolingue : première version basique nécessaire pour le développement du correcteur et du lemmatiseur*

FIGURE 3 – Feuille de route 2015-2019 pour le développement numérique de la langue occitane

3 Données de cadrage

3.1 L'occitan, langue romane et européenne

L'occitan est une langue romane proche du catalan, langue déjà bien outillée. Il serait intéressant d'étudier les possibilités de transfert technologique entre les deux langues. Il existe un traducteur automatique catalan-occitan languedocien (*Opentrad*[19]) – pouvant encore être amélioré, permettant par exemple d'envisager – avec une phase postérieure de correction manuelle – un traitement massif de corpus textuels. De même, il est envisageable et même souhaitable de s'appuyer sur les développements réalisés pour les autres langues romanes – dont les syntaxes par exemple restent finalement assez proches – et en premier lieu le français.

L'occitan est également une langue partagée par trois États (France, Espagne et Italie), la question de son développement est donc « d'intérêt » européen, avec de plus des opérateurs de missions publiques basés sur des territoires éligibles aux fonds de coopération transfrontalières. La langue occitane est également voisine au sens géographique de deux autres langues dites « minorisées » – encore que co-officielles et disposant déjà d'institutions linguistiques et de technologies de support proche de certaines langue d'État : le catalan et le basque. Sur les questions linguistiques, on a vu ces dernières années se développer des deux côtés de la frontière des actions d'échange, de mutualisation voire même de transfert de technologie : on peut mentionner le partenariat entre le *Congrès*, *Elhuyar*[20] et la société *Media.kom* – avec le soutien de l'Eurorégion Aquitaine-Euskadi – autour de la création d'un dictionnaire référentiel et orthographique occitan (le *Basic*[21]), un corpus textuel en ligne[22] ainsi que la première version de traducteur automatique occitan-français (en partenariat avec la société de presse *Vistedit* et le soutien du fonds SPEL, en cours de développement) mais également le diagnostic et la feuille de route de développement numérique de l'occitan 2015-2020 (avec la participation de représentants de l'Université du Pays Basque et de l'Université Polytechnique de Barcelone). *Le Cirdòc – médiathèque occitane* collabore depuis plusieurs années avec les centres de ressources des occitanophones italiens (vallées du Piémont) et espagnols (Val d'Aran) dans le cadre du développement de la plate-forme numérique *Occitanica*[23]. Il est également, dans le cadre de l'Eurorégion Pyrénées-Méditerranée, chef de file du forum interrégional « patrimoine et création » (réseau d'opérateurs autour du patrimoine et de la création). Enfin, on peut citer également le partenariat entre *l'InÒc Aquitaine* et *l'Institut d'Études occitanes* (Toulouse) avec le *Termcat* (Barcelone) autour de la question du développement de la terminologie en langue occitane[24].

3.2 La question des standards

Les technologies du langage utilisent des standards (graphiques, lexicaux, grammaticaux). Or, l'occitan a la particularité de ne pas avoir de variante dite « standard » mais est au contraire d'être composé de plusieurs « grandes variantes » (auvergnat, gascon, languedocien, limousin, provençal et vivaro-alpin). De même, pour ce qui concerne la graphie, Lo Congrès utilise, codifie et diffuse la graphie dite « classique[25] » par ses productions (dont un dictionnaire d'orientation pan-occitan – *lo Basic* – et un conjugateur). De par sa mission d'organisme de régulation de la langue occitane confiée par ses partenaires publics, *Lo Congrès* s'emploie à répondre aux besoins de la transmission de la langue en stabilisant les formes en graphie classique autour des grands espaces linguistiques. Cela n'écarte en rien les autres systèmes, les technologies permettant d'envisager le développement de transcripteurs graphiques et dialectaux. Il conviendra d'en étudier la faisabilité technique et budgétaire, ainsi que les besoins réels.

3.3 La recherche scientifique

L'occitan dispose d'un réseau international de chercheurs (rassemblés au sein de l'Association internationale d'études

19 *Opentrad* est une plateforme de traduction automatique basée sur le moteur au code source ouvert *Apertium* : http://www.opentrad.com
20 https://www.elhuyar.eus
21 http://locongres.org/index.php/fr/lo-congres-fr/les-chantiers/la-basic/introduction
22 http://corpus.locongres.org
23 http://www.occitanica.eu
24 Dans le cadre de la céation d'un lexique du transport touristique pour plusieurs langues latines minorisées (occitan, corse, sarde, etc.) : http://www.termcat.cat/ca/Diccionaris_En_Linia/36/Fitxes/
25 Autres graphies : mistralienne, fébusienne, escolo dóu Po, etc.

occitanes et pour certains présents au Conseil linguistique du Congrès), plusieurs départements à l'Université ainsi que deux laboratoires actifs dans les domaines des TAL :

— CLLE est un laboratoire de recherche en psychologie et en linguistique. Sa composante, CLLE-ERSS (Cognition, Langues, Langage, Ergonomie – Equipe de Recherche en Syntaxe et Sémantique, basée à l'Université Toulouse 2) a pour visée la description scientifique et la modélisation des langues naturelles (phonologie, morphologie, syntaxe, sémantique, lexique, discours). Depuis sa création, CLLE-ERSS s'engage dans la constitution et l'exploration de grands corpus langagiers, écrits ou oraux.

— *BaTelÒc* (base textuelle occitane) est un projet transversal CLLE-ERSS, sous la responsabilité de Myriam Bras, visant la construction d'une base de textes annotée en langue occitane, en partenariat avec le CNRTL, le CROM, CIEL d'ÒC, IEO/IDECO, lo Congrès Permanent de la Lenga Occitana, lo CIRDOC avec le soutien financier de la Région Midi-Pyrénées.

— CLLE_ERSS participe également avec LILPA – Université de Strasbourg, LESCALP – Université d'Amiens, LIMSI – Université Paris-Orsay au projet *Restaure*[26] (Ressources Informatisées et Traitement Automatique pour les langues régionales) ; il s'agit d'un programme ANR (2015-2018, appel à projet 2014) visant à développer des ressources et des outils de traitement automatique des langues (TAL) pour trois langues de France : le picard, l'alsacien et l'occitan.

— Le laboratoire UMR 730 *base, corpus, langage* associé au CNRS de l'Université de Nice travaille sur le corpus et exploite les données à des fins théoriques autour de quatre thématiques : dialectologie-phonologie-diachronie, logométrie et corpus politiques, médiatiques et littéraires, linguistique de l'énonciation, langage et cognition. Il est éditeur de l'importante base *Thésoc*[27] (Thesaurus occitan) : il s'agit d'un important programme comprenant une base lexicale annotée de 1,2 millions de mots avec un volet cartographique permettant de visualiser les variantes dialectales et un module morphosyntaxique (MMS).

3.4 Les opérateurs institutionnels

Il est à noter que l'occitan a trois opérateurs de missions publiques possédant une expertise technique pour la collecte, la numérisation, le traitement et l'édition numérique ainsi que de nombreuses ressources et outils :

— Le Congrès est spécialisé dans le traitement numérique des ressources linguistiques occitanes. Il possède des ressources linguistiques de référence (dictionnaires, modèles de conjugaison), son équipe professionnelle (lexicographie, développement et webmastering) et ses différents prestataires lui permettant d'assurer la chaîne complète de traitement de données : numérisation, « parsage », formatage, développement applicatif et édition numérique.

— *Le Cirdòc – médiathèque occitane* est un établissement public travaillant à la sauvegarde, à la valorisation et à la diffusion du patrimoine occitan. Pôle associé à la BNF pour la langue et la culture occitanes, *le Cirdòc* développe des actions inter-régionales autour du patrimoine et de la création occitans : la médiathèque numérique www.occitanica.eu, le développement de la numérisation des documents en partenariat avec les institutions de toutes les régions et leur diffusion dans le cadre d'Occitanica, la création d'outils de connaissance et de développement de l'occitan (Répertoire des fonds occitans, Bibliographie occitane), la conception et prêts d'expositions ou de ressources documentaires (service Question/Réponse, service aux chercheurs). Avec plus de 80 000 titres du XVIe siècle à nos jours (manuscrits, archives, livres, revues, partitions, enregistrements sonores et audiovisuels, estampes, affiches, photographies, objets, etc.), il est le grand conservatoire de la langue et culture occitane.

— *l'InÒc Aquitaine* a pour mission régionale la valorisation des ressources numériques de l'occitan, qu'il s'agisse des pratiques vivantes et des savoir-faire (Patrimoine culturel immatériel) en Aquitaine ou de fonds patrimoniaux anciens. Ethnopôle[28], il conçoit et réalise pour ce faire des projets éditoriaux en ligne (sites internet) dans le cadre d'*Aquitaine Cultures Connectées* (anciennement BnsA) : *Sondaqui*[29] (patrimoine oral et festif aquitain), *Troubadours d'Aquitaine*[30] et collabore avec la *Banque Numérique des ressources Pyrénéennes* (BNRP). Dans sa mission régionale de préservation et de valorisation du patrimoine sonore et audiovisuel occitan et de la mémoire collective en Aquitaine. Il accompagne,

26 http://lilpa.unistra.fr/fdt/projets/projets-en-cours/restaure/
27 http://thesaurus.unice.fr/
28 L'appellation ethnopôle est un label du ministère de la Culture et de la Communication attaché à une institution qui, en matière de recherche, d'information et d'action culturelle, œuvre à la fois au plan local et au niveau national. À travers cette appellation, la mission du patrimoine ethnologique entend, dans le cadre propre à chaque structure, promouvoir une réflexion de haut niveau s'inscrivant tout à la fois dans les grands axes de développement de la discipline ethnologique et dans une politique de constitution des bases d'une action culturelle concertée. Site internet : http://www.culturecommunication.gouv.fr/Politiques-ministerielles/Patrimoine-ethnologique/Ethnologie-en-region/Ethnopoles
29 http://www.sondaqui.com
30 http://www.trobar-aquitaine.org

pour ce faire, les collectivités dans les différentes étapes du processus de sauvegarde des archives (inventaire, numérisation, description, collecte). Son pôle *Langue et Société* travaille sur la terminologie (lexiques spécialisées), la toponymie et mè, e un important travail de traduction.

— Les 23 et 27 juin 2014, les Régions Aquitaine et Midi-Pyrénées ont respectivement approuvé en assemblée plénière la création de l'Office public de la langue occitane, un nouvel outil dédié à la promotion de la langue. L'objectif de ce G.I.P (Groupement d'Intérêt Public) sera d'assurer la sauvegarde et le développement de l'occitan, en travaillant à l'augmentation quantitative et qualitative du nombre de locuteurs. Il participera à la mise en œuvre d'une politique linguistique publique interrégionale. Cet outil commun, qui a pour vocation d'accueillir rapidement d'autres partenaires, comme l'État ou d'autres Régions, est en cours d'installation ; il remplacera à terme les régions membres dans leur soutien au Congrès, dont il deviendra le partenaire public privilégié.

3.5 La « communauté »

Bien que modeste, il existe une communauté numérique occitane , c'est-à-dire des particuliers participant à des projets collaboratifs et en code-source ouvert en langue occitane. Nous parlions plus haut de la communauté occitanophone de Wikipédia qui a produit près de 100 000 articles ainsi qu'une version occitane du Wikictionnaire (programme de dictionnaire collaboratif), Wikiccionari[31], possédant 30 000 entrées (ce qui au regard de la situation sociolinguistique n'est pas négligeable).

De même il existe une communauté travaillant à la localisation (traduction) de logiciels, malheureusement par faute de moyens pas toujours concertés et connus du grand public. L'association *Tot en Òc* a traduit *Firefox* et *Libre Office* en occitan, le système d'exploitation libre *Ubuntu* est traduit à 80 %. Nous pouvons citer également *Dicollecte*[32], un projet collaboratif visant à améliorer les dictionnaires orthographiques pour les logiciels libres (la version occitane comprend actuellement un peu plus de 65 000 entrées).

4 Vers un programme opérationnel

Le plan de développement numérique de la langue occitane est un programme complexe rassemblant différents acteurs et dispositifs administratifs et budgétaires. Il est indispensable d'être vigilant sur la qualité de la planification (ordre des réalisations) et de la coordination des différents acteurs[33]. Aussi, pour garantir le bon déroulement du programme, il conviendrait de s'assurer :

— au niveau de la maîtrise d'ouvrage du projet, de reconduire le Comité de pilotage politique de la feuille de route, qui pourrait s'élargir à d'autres participants (universités ? experts extérieurs ?). Espace de concertation entre des décideurs et des prescripteurs, il valide les objectifs, le calendrier et les moyens du programme.

— au niveau de la maîtrise d'œuvre, que le pilotage soit assuré par le Congrès permanent de la langue occitane. Il rassemble en son sein les différents acteurs concernés par le développement de cette feuille de route et a une vision transversale de la problématique du TAL. En tant qu'organisme de régulation de la langue occitane, il dispose également de ressources de référence en TAL indispensables le plaçant en position centrale pour la réalisation des différents objectifs. Pour des raisons de capacité et de moyens, *le Congrès* ne sera pas en mesure de porter seul le développement de la feuille de route. Cette maîtrise d'œuvre comportera donc sur les aspects techniques et opérationnels une assistance ainsi que des coproductions avec différents opérateurs : *le Cirdòc* (qui possède la surface financière nécessaire au portage et à la gestion des dossiers européens), Université de Toulouse Jean-Jaurès, *InÒc Aquitaine*, *Elhuyar* (pour les développements applicatifs) pour ne citer qu'eux.

Pour ce qui concerne la partie opérationnelle, le rapport propose un développement des ressources et outils que l'on pourrait diviser en trois catégories :

- Les ressources de base, dans lesquelles on distingue les corpus et les bases lexicales (monolingues et bilingues) ;
- Les outils intermédiaires (analyseur morphosyntaxique, analyseur syntaxique, reconnaissance d'entités nommées) ;
- Les outils finaux (correcteur orthographique, traducteur automatique).

La feuille de route et de son calendrier de réalisation d'objectifs permettent, en considérant les interdépendances entre ses différences ressources, de fixer des objectifs de réalisation.

31 http://oc.wiktionary.org
32 http://www.dicollecte.org
33 Les experts internationaux ont particulièrement insisté sur ces aspects lors de l'étude.

OBJECTIF	RESSOURCE/OUTIL NÉCESSAIRE
Corpus monolingue	Numérisation, OCR et conversion de texte à un format standard traitable par un analyseur
Corpus web	Détecteur de l'occitan Détecteur des variantes de l'occitan
Corpus parallèle	Collection de documents bilingues Mémoires de traduction (TMX)
Base lexicale monolingue	Dictionnaires monolingues au format électronique (MRD)
Base lexicale bilingue	Dictionnaires bilingues au format électronique (MRD)
Correcteurs orthographiques	Base lexicale monolingue
Analyseur morphologique (*tagger*, lemmatiseur)	Base lexicale monolingue Base grammaticale
Analyseur syntaxique	Analyseur morphologique Base grammaticale/syntaxique
Base de connaissance lexicale	Base lexicale monolingue
Traducteurs automatiques oc → fr (toutes les variantes)	Base lexicale bilingue Base grammaticale/syntaxique
Transcripteur automatique entre variantes	Base lexicale monolingue Base grammaticale/syntaxique

FIGURE 4 – Outils et ressources : interdépendances

Dans l'attente d'un programme opérationnel détaillé de développement de la feuille de route, deux travaux sont d'ores et déjà lancés : un lexique ouvert des formes fléchies d'occitan (partenariat CLLE-ERSS/Lo Congrès permanent de la lenga occitana) concernant, pour commencer, les variétés gasconne et languedocienne et un traducteur automatique occitan (gascon et languedocien)-français sur la plate-forme Opentrad en partenariat avec la fondation Elhuyar.

Conclusion

Si la langue occitane pâtit d'un retard important en TAL, elle est également en capacité de se doter dans des délais raisonnables des premières technologies de support. La feuille de route de développement numérique 2015-2019 en détaille la faisabilité dans un calendrier opérationnel. On constate également que l'occitan dispose de réels atouts, tels que le dynamisme de la recherche universitaire et des acteurs institutionnels, sa proximité avec les autres langues romanes ou encore des perspectives en termes de transferts de technologies avec les langues géographiquement voisines.

Cependant, il est indispensable, comme le souligne le Livre blanc de META-NET, de mettre en place une planification rigoureuse et de se donner les moyens de mutualiser efficacement les ressources. Cette feuille de route, largement

diffusée depuis, a permis de sensibiliser et, il faut l'espérer, de mobiliser les différents agents de transmission de la langue à des enjeux vitaux pour la langue occitane.

Il s'agit d'un programme pluriannuel rassemblant des acteurs (universités, institutions, associations) ayant leurs propres contraintes administratives, techniques et financières et répartis sur un grand espace géographique (huit régions administratives françaises, une italienne et une autonomie espagnole) qui donc nécessitera un niveau élevé de coordination.

Enfin, même si elle n'est pas chiffrée, la feuille de route représentera, on le sait, un budget conséquent. Les montants toutefois pressentis induisent en amont un important travail d'ingénierie financière, avec la participation active des collectivités, en particulier les Régions, des universités et des opérateurs institutionnels. Il s'agit d'un pré-requis indispensable pour émarger sur la nouvelle programmation 2015-2020 des crédits FEDER et les crédits de coopération interrégionale et transfrontalière.

Remerciements

Nous remercions l'ADEPFO qui a financé l'étude ayant permis la rédaction de la feuille de route pour le développement numérique occitan, la fondation Elhuyar qui en a assuré le support technique ainsi que tous les experts qui y ont porté leurs concours. Nous remercions également Marianne Vergez-Couret ainsi que les relecteurs anonymes pour leurs conseils avisés.

Références

BEC P. (1995), *La langue occitane*, Number 105, Que sais-je ? Paris.

GURRUTXAGA A. (2014). *Diagnostic et feuille de route pour le développement numérique de la langue occitane* : 2015-2019. Rapport final de la fondation Elhuyar dans le cadre de l'étude-action ADEPFO. Billère.

LEIXA, J., MAPELLI, V. et CHOUKRI, K. (2014). *Inventaire des ressources linguistiques des langues de France.* Rapport réalisé par ELDA en partenariat avec la DGLFLF-ministère de la Culture et de la Communication. Paris.

REHM, G. & USZKOREIT, H., editors. (2012). *META-NET White Paper Series: Europe's Languages in the Digital Age*. Springer, Heidelberg etc. 32 volumes on 31 European languages

SORIA C., MARIANI J., ZOLI C. 2013. *Dwarfs sitting on the giants' shoulders – how LTs for regional and minority languages can benefit from piggybacking major langages.* In M.J. Norris, E. Anonby, M-O. Junker, N. Ostler and D. Patrick (Eds.). Proceedings of the XVII FEL Conference. Carleton University, Ottawa, Canada, 1-4 October 2013, pp. 73-79

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).

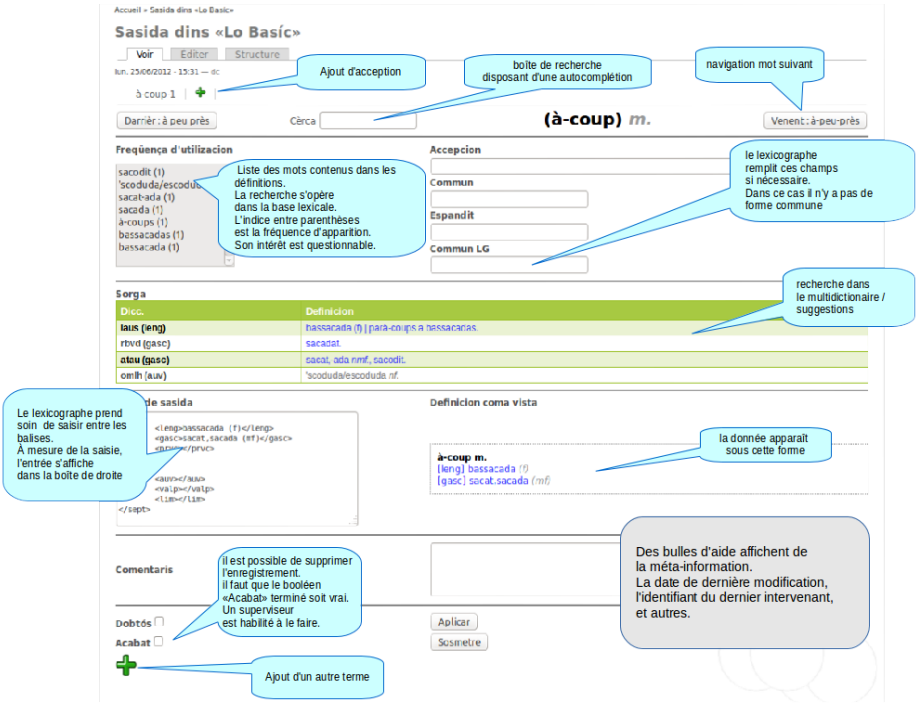


FIGURE 2 – Interface de saisie pour les lexiques (copie d'écran)

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*

Analyse morphosyntaxique de l'occitan languedocien : l'amitié entre un petit languedocien et un gros catalan

Marianne Vergez-Couret[1] Assaf Urieli[1,2]

(1) CLLE-ERSS, CNRS, Université de Toulouse 2, 5 allées Antonio Machado, 31058 TOULOUSE cedex 9
(2) Joliciel Informatique SARL, 2 avenue du Cardié, 09000 FOIX
marianne.vergez@univ-tlse2.fr, assaf.urieli@univ-tlse2.fr

Résumé. Dans cette étude, nous nous intéressons à la question de l'analyse morphosyntaxique de l'occitan. Nous utilisons Talismane, un logiciel par apprentissage supervisé, nécessitant des données annotées pour l'entraînement et optionnellement un lexique. Nous montrons dans cet article**,** qu'en l'absence de données annotées suffisantes pour l'occitan, il est possible d'obtenir de bons résultats (92%) en utilisant les données d'une langue étymologiquement proche, le catalan. Nous avons utilisé le corpus Ancora (500 000 formes) et un lexique occitan languedocien (250 000 entrées). Utiliser un corpus catalan de taille importante permet une amélioration de +3% par rapport au résultat obtenu avec le seul corpus d'entraînement occitan disponible à ce jour de 2800 formes.

Abstract.

Pos-tagging the Lengadocian dialect of Occitan: a little Lengadocian befriends a big Catalan.
In this study, we examine the question of Occitan POS-tagging. We use Talismane, a supervised machine learning NLP tool, requiring annotated data for training and optionally a lexicon. We show that, with insufficient data for Occitan, it is possible to obtain good results (92%) by using data from an etymologically close language, in this case Catalan. We used the Catalan Ancora corpus (500,000 tokens) and an Occitan Languedocien lexicon (250,000 entries). Using the larger Catalan corpus improved results by +3% with respect to the result obtained using the only Occitan training corpus available to date (2,800 tokens).

Mots-clés : traitement automatique des langues peu dotées, occitan, analyse morphosyntaxique

Keywords: natural language processing for lesser resourced languages, Occitan, POS-tagging.

1 Introduction

Les méthodes les plus couramment employées pour développer des outils de TAL sont à l'heure actuelle des méthodes par apprentissage supervisé quand des données annotées sont disponibles. Nous inscrivons nos travaux dans cette tendance pour l'analyse morphosyntaxique automatique de l'occitan. Il est donc nécessaire de rassembler des lexiques et des corpus annotés. Construire ces ressources requiert des efforts conséquents et des moyens financiers et humains qui font souvent défaut dans le cas des langues peu dotées[1] comme l'occitan.
Dans cet article, nous proposons de comparer les résultats obtenus avec un corpus d'entraînement languedocien de petite taille (2800 formes) et un corpus d'entraînement catalan de grande taille (500 000 formes). Le corpus languedocien a l'avantage d'avoir été annoté par nos soins et correspond parfaitement aux besoins d'annotation attendus. Il favorise la qualité au détriment de la quantité. Le corpus catalan que nous utilisons pour cette étude est Ancora (Taulé et al., 2008), un corpus de 500 000 formes annotées. Nous souhaitons évaluer s'il faut, dans la constitution ou l'exploitation de ressources pour l'entraînement d'un analyseur morphosyntaxique, favoriser la qualité ou la quantité des annotations. Pour ce faire, nous avons mis en place plusieurs expériences : a) entraînement avec chacun des deux corpus séparés (après une étape d'harmonisation) ; b) entraînement avec transformation superficielle du corpus catalan avec transposition en occitan des mots les plus fréquents ; c) entraînement avec combinaison des deux corpus en faisant varier le poids attribué au corpus languedocien.

[1] Le terme "langues peu dotées" pour les langues disposant de peu ou pas de ressources et d'outils linguistiques informatisées sera utilisé en opposition à "langues très dotées" pour celles qui disposent d'un grand nombre de ressources et d'outils.

En section 1, nous présentons l'occitan et le catalan ainsi que les principaux traits de ressemblance et de différence lexicale, morphologique et syntaxique des deux langues. Puis, nous présentons en section 2 les principes de fonctionnement de Talismane (Urieli, 2013) que nous avons choisi pour cette étude avant d'aborder plus généralement en section 3 quelques méthodes courantes en traitement automatique des langues peu dotées. Nous présentons, en section 4, les ressources que nous avons rassemblées pour cette étude et en section 5 les expériences que nous avons menées ainsi que les résultats que nous avons obtenus.

2 Deux langues étymologiquement proches : Occitan et Catalan

L'occitan et le catalan font partie de l'ensemble des langues romanes. Elles sont toutes deux issues de la fragmentation dialectale de l'ensemble gallo-roman méridional. Pierre Bec (1995) divise cet ensemble en quatre *complexus dialectaux :* le nord-occitan (limousin, auvergnat, vivaro-alpin), l'occitan méridional (languedocien, provençal), le gascon et le catalan. Il faut souligner que l'appartenance du catalan à cet ensemble gallo-roman est une position idéologique qui n'est pas partagée par tous. Néanmoins, P. Bec explique cela par les «extraordinaires ressemblances" que l'occitan et le catalan présentent. Toutefois, les deux langues, bien que très proches à l'origine, furent séparées politiquement et culturellement à partir du 13ème siècle comme le souligne Sibille (1996). Bec (1995) distingue le catalan de l'occitan méridional, et en particulier du languedocien, par des traits phonétiques (par ailleurs répercutés sur la graphie) et une certaine originalité du lexique. A l'heure actuelle, l'occitan et le catalan ont des situations sociolinguistiques bien différentes. Nous résumons ci-dessous quelques caractéristiques de ces deux langues.

2.1 Occitan

L'occitan est parlé dans le sud de la France et dans quelques vallées espagnoles et catalanes. Il est difficile d'estimer le nombre de locuteurs occitans mais plusieurs études permettent d'établir un chiffre aux alentours de 500 000 locuteurs. Néanmoins, un nombre plus important de locuteurs, estimés aux alentours de 1,5 millions de personnes[2], peuvent manifester un intérêt pour la langue occitane (la pratiquer et/ou la comprendre avec divers degrés de compétences). Cet intérêt pour l'occitan est appuyé par un important réseau associatif parmi lequel les écoles primaires et secondaires en immersion bilingue Calandreta, l'Institut d'Estudis Occitan (IEO), le CFPO (Centre de Formacion Professionala Occitan) qui offrent des formations en occitan pour tous les âges. L'occitan est également présent dans le système éducatif français sous forme d'options obligatoires ou facultatives du primaire à l'université.

2.2 Catalan

Le catalan est parlé en Catalogne autonome, en Catalogne Nord (Pyrénées Orientales), en Andorre, en Pays Valencien, dans les îles Baléares, dans la Frange orientale d'Aragon, dans la ville d'Alguer (en Sardaigne) et dans la région du Carxe en Murcie, regroupant 9,6 millions de locuteurs sur un total de 13,5 millions d'habitants (Almarcha París et Baylac Ferrer, 2012). La pratique courante de la langue catalane dans l'enseignement, les médias, la politique et les entreprises place le catalan dans une position intermédiaire entre langues normalisées (comme le castillan) et langues minorisées (comme l'occitan), bien qu'Almarcha París et Baylac Ferrer (2012) signalent des disparités de cette pratique selon les régions concernées. La quantité de données linguistiques disponibles pour le catalan qui occupe une place de choix sur internet (Serra Serra, 2012) est également un indice fort de cette position (Boleda *et al*, 2009). Selon nous, le catalan est dans l'ensemble des langues romanes, la langue la plus proche de l'occitan. Nous souhaitons donc mettre en place une méthode permettant de tirer parti au mieux des ressources linguistiques disponibles pour le catalan (dans notre cas le corpus Ancora[3]) pour une tâche d'analyse morphosyntaxique de la variante languedocienne de l'occitan. Nous allons présenter dans la section suivante des traits similaires et différents des deux langues.

2.3 Ressemblances et différences lexicales, morphologiques, syntaxiques

Nous allons décrire dans les sections suivantes quelques ressemblances et différences majeures de la variante languedocienne de l'occitan et du catalan standard sans toutefois viser l'exhaustivité.

[2] D'après une étude socio-linguistique réalisée en Midi-Pyrénées en 2010 (http://www.midipyrenees.fr/IMG/pdf/EnqueteOccitan.pdf)

[3] De nombreux outils de TAL (FreeLing) et des lexiques (Apertium) sont également disponibles librement pour le catalan et pourront être exploités pour d'autres expériences.

2.3.1 Caractéristiques phonétiques et lexicales

Les principales distinctions phonétiques en languedocien et en catalan sont décrites dans (Bec, 1995). Elles sont le résultat d'évolutions divergentes qui ont conduit à des distinctions phonétiques et également graphiques.

Latin	Clave	Cantare	Capra	Causa	Lingua
Catalan	Clau	Cantar	Cabra	Cosa	Llengua
Occitan	Clau	Cantar	Cabra	Causa	Lenga

Tableau 1. Mots occitans et catalans d'origine latine

Par exemple, l'occitan a conservé la diphtongue *au* et pas le catalan *causa/cosa*. Le couple *llengua/lenga* illustre la palatalisation du l- initial en catalan. Bec signale également une certaine originalité du lexique du catalan, par exemple pour les mots proches du castillan comme *molt.*

2.3.2 Ressemblances et différences morphologiques

Tandis que le masculin pluriel et le féminin singulier réguliers des adjectifs catalans et occitans sont similaires, les tableaux ci-dessous montrent la différence pour le féminin pluriel.

	Masculin/Singulier	Masculin/Pluriel	Féminin/Singulier	Féminin/Pluriel
Catalan	Madur	Madurs	Madura	Madur**es**
Occitan	Madur	Madurs	Madura	Madur**as**

Tableau 2. Adjectif *Madur (mûr)* en catalan et en occitan

De nombreuses différences peuvent également être relevées dans les flexions verbales de l'occitan et du catalan.

Catalan	Cant**o**	Cant**es**	Canta	Cant**em**	Cant**eu**	Cant**en**
Occitan	Cant**i**	Cant**as**	Canta	Cant**am**	Cant**atz**	Cant**an**

Tableau 3. Verbe cantar (chanter) présent de l'indicatif en catalan et en occitan

2.3.3 Différences syntaxiques

Nous proposons dans cette section de dégager quelques caractéristiques de l'occitan et du catalan à partir des exemples suivants en languedocien (pour les parties a) et en catalan (pour les parties b) de la *Parabole de l'Enfant prodigue* repris de Bec (1995) en nous focalisant sur celles qui ont un impact sur l'ordre et la distribution des mots dans les deux langues.

La négation ne se marque pas de la même façon en catalan et en languedocien. En catalan, la marque de la négation se place avant le verbe, cf «només tenia" en 1 et en «no va tenir" en 2 tandis qu'en languedocien la ou les marques de négation seront exprimées après le verbe «aviá pas que" en 1 et «aguèt pas mei" en 2.

1. *a) Un òme aviá **pas que** dos dròlles.*
   *b) Un home **només** tenia dos fills.*
   'Un homme n'avait que deux fils.'

2. *a) **Aguèt pas mei** de lèit per dormir la nuèit ni de fuòc per **se calfar** quand aviá freg.*
   *b) Ja **no va tenir** llit per a dormir a la nit ni foc per a **escalfar-se** quan tenia fred.*
   'Il n'eut plus de lit pour dormir la nuit ni de feu pour se chauffer quand il avait froid.'

Il existe une différence majeure en catalan et en languedocien concernant la structure *anar* (aller) + verbe à l'infinitif qui en catalan permet d'exprimer le prétérit comme en 3 : «va dir" et en languedocien permet d'exprimer le futur proche.

3. *a) Lo plus jove **diguèt a son** paire :*
   *b) El més jove **va dir al seu** pare :*
   'Le plus jeune dit à son père :'

Mais au final, cette différence interprétative n'a aucune incidence sur l'annotation morphosyntaxique que nous proposons qui sera la même en catalan et en languedocien :

va anar Vc-Pri-P3-sg (Verbe conjugué au présent de l'indicatif, troisième personne du singulier)
cantar cantar Vi (Verbe à l'infinitif)

La position des pronoms en catalan et en languedocien est un problème complexe et sujet à beaucoup de variations. Toutefois, dans les phrases conjuguées, les pronoms se placent avant le verbe, par exemple en 4 « *li balhava* » en languedocien et « *li donava* » en catalan. L'ordre des pronoms peut changer en catalan et en languedocien : « *me la compro* » vs. « *la me crompi* » mais ce phénomène aura probablement peu de conséquences pour notre tâche d'annotation morphosyntaxique étant donné que les deux pronoms auront la même catégorie grammaticale (pronom clitique). En revanche, en catalan standard, le pronom apparaît toujours après le verbe lorsqu'il est à l'infinitif, par exemple « *escalfar-se* » (« *se calfar* » en languedocien) en 5 et « *anar-me'n* » en 6, ce qui est impossible en languedocien.

4. *a) Mas degun* ***li balhava*** *pas res.*
   b) Pero ningú no ***li donava*** *res.*
   'Mais personne ne lui donnait rien.'

5. *a)* ***Aguèt pas mei*** *de lèit per dormir la nuèit ni de fuòc per* ***se calfar*** *quand aviá freg.*
   b) Ja ***no va tenir*** *llit per a dormir a la nit ni foc per a* ***escalfar-se*** *quan tenia fred.*
   'Il n'eut plus de lit pour dormir la nuit ni de feu pour se chauffer quand il avait froid.'

6. *a) Es ora per ièu de me governar sol e d'aver argent ;* ***me*** *cal poder partir e véser de païs.*
   b) Ja és hora que sigui el meu propi amo i que tingui diners ; cal que pugui ***anar-me'n*** *i veure món.*
   'Il est temps pour moi d'être indépendant et de gagner de l'argent ; il faut que je parte et que je voie du pays.'

Il existe en occitan et en catalan plusieurs façons de marquer la possession. En catalan standard, on retrouve principalement les formes composées (*el meu, la meva, els meus, les meves, el nostre, la nostra, els nostres, les nostres*…) bien qu'il existe également des formes simples (*ma, ta, sa, vostre, nostre, son*). En 3, une forme simple en languedocien « *a son paire* » est traduite par une forme composée « *al (a+el) seu pare* » en catalan tandis qu'en 7 on trouve une forme composée dans les deux cas : « *lo vòstre ben* » et « *el vostre bé* » . Il est également possible de trouver un adjectif possessif après le nom : tandis que ce sera un phénomène rare en catalan illustré en 8, « *fill meu* », il permet avec plus de liberté de marquer une insistance en languedocien : « *l'amic mieu* » ; « *la lenga nòstra* » .

7. *a) Despartissètz* ***lo vòstre*** *ben e donatz-me çò que devi aver.*
   b) Partiu ***el vostre*** *bé i doneu-me el que m'escaigui.*
   'Partagez votre bien et donnez-moi ce qui me revient.'

8. *a) « O* ***mon filh*** *» diguèt lo paire, « coma voldràs tu ; siás un marrit e seràs castigat »*
   b) « Ai, ***fill meu*** *» , va dir el pare, « com vulguis ; ets dolent i seràs castigat »*
   'Ah, mon fils ! » dit le père, « comme tu voudras, tu es méchant et tu seras punis'

Le partitif comme dans l'exemple 9 en languedocien « *qu'an* ***de*** *pan e* ***de*** *vin,* ***d'****uòus e* ***de*** *formatge* » n'existe pas en catalan : « *que tenen pa i vi, ous i formatge* » .

9. *a) « Enlà, l'ostal del paire es plen de vailets qu'an* ***de*** *pan e* ***de*** *vin,* ***d'****uòus e* ***de*** *formatge tant que vòlon. »*
   b) « Allà a-baix, la casa del meu pare és plena de mossos que tenen pa i vi, ous i formatge, tant com en volen »
   '« Là-bas, la maison du père est pleine de serviteurs qui ont du pain, du vin, des œufs et du formage autant qu'ils en veulent. »'

Pour conclure, le Tableau 4 synthétise ces différences ci-dessous.

	Catalan standard	**Occitan**
Place de la négation simple	Avant le verbe	Après le verbe
Anar+verbe	Prétérit	Futur proche
Place des clitiques pour les verbes à l'infinitif	Après le verbe	Avant le verbe
Possessifs	Composés	Simples ou composés
Partitif	Non	Oui

Tableau 4. Synthèse des différences entre le catalan standard et l'occitan languedocien

Néanmoins, nous comptons sur tout ce qui rapproche les deux langues (pas de pronom clitique sujet, place de l'adjectif préférée derrière le nom, …) et posons l'hypothèse que la probabilité des séquences d'étiquettes grammaticales sera suffisamment similaire pour améliorer nos résultats en utilisant une ressource catalane pour entraîner un analyseur morphosyntaxique de l'occitan languedocien avec Talismane que nous présentons dans la section suivante.

3 Talismane

3.1 Fonctionnement

Dans cette étude, nous avons entrainé l'analyseur morphosyntaxique Talismane (Urieli, 2013), distribué sous une licence libre[4], sur des corpus d'entraînement occitans et catalans. Talismane a déjà été appliqué à l'anglais et au français, avec une exactitude de 97 % (Urieli, 2014), et à l'occitan avec une exactitude de 89 % (Vergez-Couret et Urieli, 2014). Talismane permet d'intégrer un lexique à la fois sous forme de descripteurs et de règles. En tant que descripteur, le lexique nous permet de dire « si le mot X existe dans le lexique en tant que nom commun, alors il est plus probable qu'il soit réellement un nom commun ». Cette information est incorporée dans le modèle statistique pendant l'entraînement, avec d'autres descripteurs listés ci-dessous. En tant que règle, le lexique nous permet de contourner les choix du modèle statistique pendant l'analyse, soit en imposant ou en interdisant le choix d'une certaine étiquette. Par exemple, on peut définir la règle suivant laquelle « le mot X ne peut être étiqueté comme préposition que s'il est listé comme préposition dans le lexique ».
Nous avons effectué une recherche de plusieurs combinaisons traditionnellement utilisées en apprentissage automatique et sélectionné la meilleure configuration pour l'occitan avec un classifieur SVM linéaire avec $\varepsilon = 0{,}1$ et $C = 0{,}5$.

3.1.1 Descripteurs

Nous utilisons le même jeu de descripteurs pour l'occitan que pour le français et l'anglais. Ces descripteurs ont été choisis en premier lieu en suivant l'intuition qu'ils indiqueront des distributions particulières d'étiquettes. Nous présentons ci-après ceux qui ont été validés et retenus après une évaluation empirique. Pour analyser un token T_i, on examine les tokens en position T_{i-2}, T_{i-1}, T_i, T_{i+1}, T_{i+2}. Les descripteurs de base comprennent pour chacun de ces tokens : **W** la forme lexicale exacte, **P** l'étiquette attribuée au token T_j (si $j < i$) ou les étiquettes trouvées dans le lexique (si $j \geq i$), **L** le lemme trouvé dans le lexique pour chaque étiquette donnée, **U** si le token est inconnu dans le lexique, $\mathbf{Sfx}_n$ les *n* dernières lettres de la forme (*n* de 2 à 5), **1st** si le token est le premier de la phrase, **Last** si le token est le dernier de la phrase. Ces briques de base sont aussi combinées en bigrammes et trigrammes. Ainsi, par exemple, $\mathbf{P}_{i-1}$ estime la distribution des probabilités pour l'étiquette du token actuel étant donné l'étiquette attribué au token précédent. $\mathbf{P}_{i-2}\mathbf{P}_{i-1}$ estime la distribution des probabilités pour cette même étiquette étant donné les étiquettes attribuées aux deux tokens précédents.

3.1.2 Règles

Les règles suivantes ont été définies autour des étiquettes des classes fermées (i.e. des catégories fonctionnelles non productives) et des classes ouvertes (i.e. des catégories lexicales productives).

- Classes fermées : l'analyseur peut attribuer une étiquette de classe fermée (e.g. prépositions, conjonctions, pronoms, …) uniquement si le token actuel est listé sous cette étiquette dans le lexique. Cette règle nous empêche, par exemple, d'inventer de nouvelles prépositions.

[4] http://redac.univ-tlse2.fr/talismane.html

- Classes ouvertes : l'analyseur ne peut pas attribuer une étiquette de classe ouverte (e.g. nom commun, adjectif, …) si le token en question est listé uniquement dans les lexiques des classes fermées. Cette règle nous empêche, par exemple, d'attribuer l'étiquette « nom commun » au token *lo* (« le » en français).

- Des règles basées sur les expressions régulières pour systématiquement attribuer les étiquettes des nombres cardinaux et de la ponctuation.

4 TA des langues peu dotées

4.1 Talismane pour l'occitan

Dans Vergez-Couret et Urieli (2014), un premier modèle de Talismane a été entraîné avec un corpus d'entraînement de 2 500 mots, un lexique de 225 386 entrées et plusieurs petits corpus d'évaluation d'environ 700 mots chacun avec pour objectif a) la création d'un premier modèle d'analyse morphosyntaxique pour l'occitan ; b) l'évaluation des performances de Talismane entraîné avec un corpus et un lexique du même dialecte, en l'occurrence languedocien, sur des corpus représentant une certaine variété dialectale ; c) l'évaluation du gain pour la création du modèle des corpus annotés *vs.* des lexiques afin de déterminer quel est l'effort le plus important pour la constitution de ressources des langues peu dotées.

4.2 Constitution des ressources : gain des corpus annotés et des lexiques

Les résultats de Vergez-Couret et Urieli (2014) tendent à montrer l'importance des lexiques, et notamment des lexiques des classes fermées (adverbes (quantifieurs, négatifs, exclamatifs, interrogatifs), déterminants, prépositions, pronoms, adjectifs possessifs). Cette conclusion va dans le sens de celle de Garrette et al. (2013) qui ont accompli une expérience où un temps limité de 4h était donné à des annotateurs pour annoter soit du corpus, soit des lexiques (construits à partir des mots les plus fréquents de corpus non annotés) pour deux langues peu dotées. Ils concluent que le gain le plus important est obtenu avec la création des lexiques.
Or s'il est toujours possible d'augmenter la quantité de corpus, cela est moins vrai pour les lexiques. Dans cet article, nous souhaitons donc nous focaliser sur le seul dialecte languedocien avec l'objectif d'obtenir le meilleur résultat possible en ne créant pas ou peu de nouvelles ressources annotées et donc en utilisant et en adaptant des ressources d'une langue étymologiquement proche.

4.3 Utilisation de ressources de langues étymologiquement proches

Dans le cas des langues peu dotées en ressources linguistiques et de TAL, des méthodes basées sur l'adaptation d'analyseurs morphosyntaxiques des langues très dotées, généralement étymologiquement proches, ont récemment été développées. Täckström et al. (2013) utilise une approche semi-supervisée basée sur un bitexte qui aligne une langue peu dotée et une langue très dotée et ont obtenu un gain significatif. Scherrer et Sagot (2013) utilisent une approche visant à identifier des cognats lexicaux entre une langue peu dotée et une langue très dotée étymologiquement proche pour récupérer la catégorie grammaticale du mot dans la langue très dotée et améliorer l'annotation de la langue peu dotée. Cette approche a l'avantage de ne nécessiter aucune ressource annotée pour la langue peu dotée. Dans le cas de l'occitan, nous ne disposons pas de bitextes pour l'occitan et une autre langue très dotée, ce qui élimine la première méthode. En revanche, nous disposons d'un lexique avec une assez bonne couverture (que nous présentons section 4.1) et des corpus annotés et déjà exploités dans Vergez-Couret et Urieli (2014), ce qui exclut finalement la deuxième méthode. Dans cet article, notre objectif est d'améliorer les résultats obtenus dans Vergez-Couret et Urieli (2014) en augmentant les corpus et en comparant les apports d'un petit corpus annoté en languedocien correspondant exactement au besoin attendu (qualité des annotations) et un grand corpus catalan (quantité des annotations) et en expérimentant plusieurs approches de transposition et de combinaison.

5 Ressources

Nous allons présenter dans cette section toutes les ressources qui ont été rassemblées, nécessaires à l'entraînement de Talismane et à la mise en œuvre de nos expériences.

5.1 Lexique et jeu d'étiquettes

La version du lexique que nous présentons ici est une version étendue de 50 000 formes (principalement de classes lexicales (adjectif, verbe, nom)) de la version utilisée dans Vergez-Couret & Urieli (2014). Ce lexique concerne uniquement la variante languedocienne de l'occitan. Il a principalement été construit avec une ressource disponible au format numérique : le dictionnaire Français/Occitan languedocien de C. Laus (2005). Les noms propres ont été extraits des lexiques Apertium (Armentano-Oller & Forcada, 2006). Des listes de formes fléchies ont été rassemblées à partir du conjugueur mis à disposition par le *Congrès permanent de la lenga occitana*. Enfin, un script a permis de générer les formes fléchies des adjectifs, des noms et des participes passés ainsi que les formes élidées et contractées des prépositions et des déterminants. Le nombre d'entrées de chaque catégorie principale est disponible dans le Tableau 5.

Etiquette	Description	Taille
A	Adjectif (général)	29656
A$	Adjectif (possessif)	85
Adv	Adverbe (général)	762
Adv$	Adverbe (négatif, quantifieur, exclamatif et interrogatif)	58
Cc	Conjonction de coordination	8
Cs	Conjonction de subordination	150
Det	Déterminant	127
Card	Cardinal	42
Cli	Pronom clitique	17
CliRef	Pronom réfléchi	17
Inj	Interjection	130
Nc	Nom commun	53449
Np	Nom propre	4609
Pct	Ponctuation	15
Pp	Participe présent	4554
Pr	Préposition	521
Prel	Pronom relatif	37
Pro	Pronom non clitique	81
Ps	Participe passé	18089
PrepDet	Préposition et déterminant amalgamé	499
Vc	Verbe conjugué	160549
Vi	Verbe à l'infinitif	5822
Z	Consonnes de liaison	3
	Total	279280

Tableau 5. Nombre de formes fléchies du lexique

5.2 Corpus d'entraînement

Pour cette étude, deux corpus d'entraînement seront utilisés : un corpus de petite taille en languedocien que nous avons annoté et un corpus de grande taille en catalan, le corpus Ancora (Taulé et al., 2008).

Le corpus languedocien (Occitan-Train) est composé d'un extrait d'une œuvre littéraire *E la barta floriguèt* d'Enric Molin et d'un article de wikipedia occitan. Ce corpus contient 3000 formes annotées manuellement avec le lemme, la catégorie grammaticale et des informations morphosyntaxiques (genre, nombre, personne, temps et aspect). Les 1000 premières formes ont été séparément annotées par trois annotateurs qui ont ensuite confronté leurs annotations pour décider d'une annotation commune et répertorier les décisions dans un manuel d'annotation. Puis 1500 mots ont été annotés par deux annotateurs qui ont également confronté leurs annotations pour décider d'une annotation commune et améliorer le manuel d'annotation. Enfin, les 500 derniers mots n'ont été annotés que par un seul annotateur.

Le corpus catalan (Catalan-Train) est une adaptation avec notre jeu d'étiquettes du corpus Ancora. Il contient 500 000 formes principalement de textes journalistiques. Le corpus Ancora est un corpus annoté à plusieurs niveaux (morphosyntaxique, syntaxique et sémantique) et contient en particulier toutes les annotations qui nous intéressent : le lemme, la catégorie grammaticale et les informations morphosyntaxiques.

5.3 Corpus d'évaluation

Pour cette étude, nous avons un corpus d'évaluation contenant des extraits de deux œuvres littéraires *Los crocants de Roergue* de Ferran Delèris (700 formes) et *Dels camins bartassièrs* de Marceu Esquieu (500). Le premier corpus a été annoté par deux annotateurs qui se sont mis d'accord sur une annotation commune après avoir confronté leurs annotations tandis que le second n'a été annoté que par un seul annotateur.

Bien que les informations morphosyntaxiques soient disponibles dans tous nos corpus (ex. genre, nombre), elles ne sont employées ni pour l'entraînement, ni pour l'évaluation. La tâche d'étiquetage pour la catégorie principale est déjà difficile pour une langue peu dotée et une fois que l'étiquette principale déterminée, les informations morphosyntaxiques sont souvent disponibles dans le lexique.

5.4 Adaptation des ressources

Une des difficultés rencontrées avec des données extraites de plusieurs sources et de plusieurs langues fut le manque de consistance des annotations entre les corpus d'entraînement et le corpus d'évaluation, les jeux d'étiquettes et les règles d'annotation pouvant être légèrement différents d'un corpus à l'autre.

De ce fait, nous avons dû procéder à des harmonisations. Une première harmonisation a été de choisir une étiquette plus générale qui couvre les annotations des deux corpus. Nous avons ainsi transformé les étiquettes CLI (pronom clitique), CLIREF (pronom réfléchi) en PRO (pronom) dans les corpus et les lexiques étant donné que les deux distinctions précédentes n'étaient pas annotées dans le corpus Ancora. Ensuite, nous avons ajouté une distinction au corpus catalan entre les adverbes de classes ouvertes (par exemple les adverbes en *–ment*) et les adverbes de classes fermées (négatifs, quantifieurs, …) sur la base d'une liste.

En plus des différences de formation et d'emploi des possessifs que nous avons brièvement abordés dans la section 1.3.3, les normes d'annotation des possessifs varient entre le corpus d'entraînement catalan et le corpus d'entraînement languedocien. Les formes possessives complexes ont été segmentées dans le corpus languedocien (« la sia » est annotée en deux formes : « la » est annotée comme déterminant défini et « sia » est annoté comme adjectif possessif) et pas dans le corpus catalan (« la_seva » est dans son ensemble annotée comme déterminant possessif). Dans ce cas, harmoniser nécessiterait une intervention au niveau de la segmentation. Nous n'avons pas effectué cette harmonisation et savons que l'annotation des adjectifs possessifs risque de ne pas être faite selon nos besoins.

5.5 Transposition du corpus catalan

Pour cette étude, nous avons adopté une méthode visant à transposer les mots plus fréquents du corpus catalan en occitan. Ce type de méthode est généralement employée pour transposer les mots les plus fréquents dans les textes d'une langue peu dotée vers leurs équivalents dans une langue très dotée étymologiquement proche (Bernhard et Ligozat, 2013 ; Vergez-Couret, 2013). Puis, l'analyseur morphosyntaxique de la langue très dotée est utilisé pour annoter la langue peu dotée. Dans cette expérience, nous faisons la transposition de la langue très dotée vers la langue peu dotée pour améliorer la couverture du lexique lors de l'entraînement et dans le but final d'améliorer les résultats lors de l'analyse. L'avantage d'effectuer la transposition dans ce sens est de pouvoir s'appuyer sur les catégories morphosyntaxiques pour transposer correctement les homographes. Nous avons construit un lexique bilingue languedocien/catalan en prenant les 250 mots catalans les plus fréquents du corpus Ancora que nous avons manuellement traduit en occitan languedocien. Nous obtenons un lexique de 150 paires de conversion (100 formes étant identiques entre les deux langues), principalement des mots grammaticaux. Les paires contenues dans la liste, une fois automatiquement transposées du catalan vers l'occitan dans Catalan-Train couvrent 93 243 changements, soit environ 19% des formes du corpus.

Les tailles des corpus et la couverture des lexiques sont synthétisés dans les tableaux ci-dessous :

Corpus	*Occitan-Train*	*Catalan-Train*	*Catalan-Train Transposé*	*Occitan-Eval*
Taille	2 840	488 389	488 389	1 214
Taille totale (sans la ponctuation) *% de formes inconnues dans le lexique*	2 368 *2,4 %*	435 814 *39,4 %*	435 814 *30,7 %*	1 018 *9,3 %*
Classes ouvertes *% de formes inconnues dans le lexique*	1 282 *4,3 %*	225 487 *62,8 %*	225 487 *51,8 %*	543 *16,9 %*
Classes fermées *% de formes inconnues dans le lexique*	1 086 *0,1 %*	205 958 *12,5 %*	205 958 *6,0 %*	475 *0,6 %*

Tableau 6. Corpus d'entraînement et d'évaluation

6 Expériences et résultats

Les ressources ont été rassemblées pour répondre aux questions suivantes :
1) Faut-il favoriser des annotations de qualité (corpus languedocien) ou des annotations en quantité (corpus catalan) ?
2) Est-il utile de faire « ressembler » le texte catalan à de l'occitan ?
3) Peut-on améliorer la qualité des résultats en combinant les deux corpus et sous quelles conditions ?

6.1 Entraînement avec un petit languedocien et un gros catalan

La première expérience vise tout simplement à entraîner Talismane avec nos corpus d'entraînement languedocien et catalan séparément : Occitan-Train (3 000 formes) ; un extrait de 3 000 formes de Catalan-Train et Catalan-Train.

Corpus d'entraînement	Occitan-Train	Catalan-Train (3 000 formes)	Catalan-Train
Exactitude	89,04	86,00	90,11

Tableau 7. Résultats

A taille égale, les annotations en languedocien permettent sans grande surprise une amélioration significative (p-valeur < 0,005, test de McNemar) par rapport aux annotations en catalan. Utiliser un gros corpus catalan (90,5 %) permet une petite amélioration non significative (p-valeur >0,1, test de McNemar) comparé au petit corpus languedocien (89 %). Nous avons souhaité voir s'il est encore possible de faire une amélioration significative de la baseline (Occitan-Train) a) en modifiant le corpus catalan (transposition en occitan des mots les plus fréquents) et b) en combinant les deux corpus languedocien et catalan.

6.2 Transposition des mots les plus fréquents

Même si les gains obtenus avec le corpus catalan ne sont pas significatifs, nous avons constaté l'impact très positif du lexique lors de l'entraînement : en effet, sans lexique, les scores baissent à 53,77%. Selon nous, le lexique fonctionne dans ce cas grâce au grand nombre de cognats (homographes entre le catalan et l'occitan qui partagent la même étiquette morphosyntaxique). Ainsi, lors de l'entraînement, Talismane « apprend » que si un mot se trouve dans le lexique, il a une forte probabilité d'être employé dans le corpus d'analyse avec la même étiquette. L'idée sous-tendant la présente expérience revient alors à faire « ressembler » encore plus le corpus catalan à de l'occitan (et à améliorer la couverture du lexique) en transposant en occitan les mots les plus fréquents du corpus catalan. 150 couples ont été automatiquement transposés du catalan vers l'occitan (sans vérification manuelle), concernant environ 19 % du corpus (cf. section 4.5) permettant de passer de 39,4 à 30,7 % de formes inconnues dans le lexique (cf. Tableau 6).

Corpus d'entraînement	Catalan-Train	Catalan-Train (transposé)
Exactitude	90,69	91,10

Tableau 8. Résultats avec transposition des mots les plus fréquents

La transposition permet une petite amélioration significative (p-valeur < 0,05, test de McNemar) sur l'intégralité du corpus Catalan-Train. Mais surtout, cela permet pour la première fois une amélioration significative de +2,06% par rapport à la baseline (p-valeur < 0,01, test de McNemar). Il serait intéressant d'augmenter le nombre de transposition des

mots les plus fréquents mais également des classes lexicales, par exemple en utilisant des méthodes d'appariement de cognats (cf. conclusion). La prochaine expérience que nous présentons vise à combiner les deux corpus d'entraînement catalan (version transposée) et languedocien.

6.3 Combinaison du corpus catalan et du corpus occitan languedocien (avec pondération)

Nous avons créé un corpus d'entraînement unique afin de joindre qualité et quantité à partir des deux corpus catalan et occitan languedocien. Puis nous avons testé plusieurs configurations faisant varier le poids du corpus occitan languedocien (en dupliquant *x* fois le corpus). Les 3 000 formes annotées en languedocien se trouvent noyées parmi les 500 000 formes annotées en catalan et la combinaison des deux corpus ne permet pas de gain significatif. Mais donner plus de poids au corpus languedocien permet d'améliorer sensiblement les résultats jusqu'à notre meilleur résultat de 92,26 avec un poids de 200 attribué au corpus languedocien, donc un gain de +1,16% par rapport au corpus catalan seul transposé (*p*-valeur = 0,0007, test binomial) et +3,22% par rapport à la baseline (*p*-valeur < 0,005, test de McNemar)..

Corpus d'entraînement	Catalan-Train + Occitan-Train ×1 (≈ 3 000)	Catalan-Train + Occitan-Train ×25 (≈ 75 000)	Catalan-Train + Occitan-Train ×50 (≈ 150 000)	Catalan-Train + Occitan-Train ×100 (≈ 300 000)	Catalan-Train + Occitan-Train ×200 (≈ 600 000)
Exactitude	91,26	91,68	92,17	92,00	92,26

Tableau 9. Résultats selon le poids attribué au corpus languedocien

7 Conclusion

Dans cet article, nous souhaitions montrer qu'il est possible d'obtenir de meilleurs résultats que Vergez-Couret et Urieli (2014) pour l'analyse morphosyntaxique de l'occitan en employant une ressource d'une langue bien dotée et étymologiquement proche, le catalan. En effet, nous avons amélioré les résultats de 3,22 % en atteignant 92,26 % en combinant une ressource catalane de grande taille, Ancora, et une petite ressource annotée en occitan languedocien et en attribuant un poids plus important au corpus languedocien. Le résultat le plus intéressant de notre point de vue est d'avoir montré par ce biais que même un petit corpus (1/200ème du corpus catalan), plus proche de la variante à évaluer, peut améliorer les scores de façon significative, ce qui est très prometteur pour l'analyse inter-variante.

Par ailleurs, nous avons testé des méthodes visant à « faire ressembler » superficiellement le corpus catalan à l'occitan en remplaçant les 250 formes lexicales les plus fréquentes par leur traduction. Les résultats sont encourageants pour 19 % de transformations sur le corpus catalan, essentiellement des catégories grammaticales. Il serait donc intéressant d'étendre la couverture des transformations aux catégories lexicales pour voir si plus de régularités de formes permettent un gain significatif. Pour ce faire, nous pensons nous inspirer des méthodes de similarité lexicale présentées dans Scherrer et Sagot (2013).

Pour cette expérience, nous avons peu évalué l'effet du genre dans les corpus d'entraînement et lors de l'évaluation, bien qu'il a été montré que cela peut avoir un effet marqué (Candito et Seddah, 2012). Notre objectif est en premier lieu d'annoter les textes littéraires de la base BaTelòc (Bras et Thomas, 2011) (dont est extrait le corpus d'évaluation), mais dès que BaTelòc va s'ouvrir à d'autres types de textes, il sera intéressant d'augmenter la taille et la variété de nos ressources d'entraînement et d'évaluation. Afin de mieux comprendre les processus d'analyse et améliorer les résultats, on serait tenté d'analyser les erreurs dans nos corpus d'évaluation. Mais cela risquerait d'induire une sur-adéquation du modèle à moins de les passer en ressources d'entraînement. Ceci définit une méthodologie de travail dans laquelle les données d'évaluation de chaque étude deviennent des ressources de développement pour l'étude suivante. A terme, un gros occitan viendra-t-il renforcer l'amitié catalano-occitane ? *A la fortuna !*

Remerciements

Nous remercions les trois relecteurs anonymes du comité scientifique de TALaRE ainsi qu'Estel Llansana et Nabil Hathout pour leurs conseils qui ont permis d'améliorer la qualité de l'article. Les travaux décrits dans cet article ont bénéficié du soutien de l'ANR (projet RESTAURE - convention ANR-14-CE24-0003-01).

Références

ALMARCHA PARIS M., BAYLAC FERRER, A. (2007). La langue des Pays Catalans. *Langues et Cité :bulletin de l'observation des pratiques linguistiques,* 21, 2.

ARMENTANO-OLLER, C., FORCADA M.-L. (2006). Open-source machine translation between small languages: Catalan and Aranese Occitan. *In Strategies for developing machine translation for minority languages (5th SALTMIL workshop on Minority Languages organized in conjunction with LREC 2006)*, pp. 51-54.

BEC P. (1995). *La langue occitane*. Number 1059. Paris : Que sais-je ?

BERNHARD, D., LIGOZAT A.-L. (2013). Es esch fàscht wie Ditsch, oder net? Étiquetage morpho-syntaxique de l'alsacien en passant par l'allemand. *In Actes de TALARE 2013 : Traitement Automatique des Langues Régionales de France et d'Europe*, pp. 209-220.

BOLEDA G., CUADROS, M., ESPANA-BONET C., MELERO, M., PADRO, L., QUIXAL, M. RODIRGUEZ, C. (2007). Primera Jornada del Procesamiento Computacional del Catalán. *Processamiento del Lenguaje Natural, núm,* 43, 387-388.

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 IXème Congrès International de l'AIEO, Aache, Shaker.

CANDITO M. et SEDDAH D. (2012). Le corpus sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. *In TALN 2012-19e conférence sur le Traitement Automatique des Langues Naturelles.*

GARETTE, D., MIELENS J., BALDRIDGE J. (2013). Real-word semi-supervised learning of pos-taggers for low-resource languages. *In Actes dela conférence de l'Association for computational linguistics (ACL)*, pp. 583-592.

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

SCHERRER Y., SAGOT B. (2013). Lexicon induction and part-of-speech tagging of non-resourced and tools for closely related languages and language variants. Actes de *RANLP Workshop on Adaptation of language resources and tools for closely related languages and language variants*, 30-39.

SERRA SERRA J. (2007). Le catalan sur internet. *Langues et Cité :bulletin de l'observation des pratiques linguistiques,* 21, 11.

SIBILLE J. (1996). Lo gascon dialecte occitan o lenga a part entièra : Es que la question a un sens ? Elements de responsa a las teorias de Jan Lafita. *Estudis Occitans,* 20, 38-40.

SIBILLE J. (2007). L'occitan, qu'es aquò ?. *Langues et Cité :bulletin de l'observation des pratiques linguistiques,* 10, 2.

TACKSTROM O., DAS D., PETROV S., MCDONALD R., NIVRE J. (2013). Token and type constraints for cross-lingual part-of-speech tagging. Actes de *Transactions of the Association for Computational Linguistics*, 1-12.

TAULE M., MARTI M.A., RECASENS M. (2008). Ancora: Multilevel Annotated Corpora for Catalan and Spanish. Actes de *6th International Conference on Language Resources and Evaluation*, 96-101.

URIELI A. (2013). *Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. Thèse de doctorat, Université de Toulouse 2 Le Mirail.

URIELI A. (2014). Améliorer l'étiquetage de « que » par les descripteurs ciblés et les règles. *In Actes de la 21ème conférence sur le Traitement Automatique des Langues Naturelles;* 56-66

VERGEZ-COURET M. (2013). Tagging Occitan using French and Castillan Tree Tagger. *In Actes de Less Resources Languages, new technologies, new challenges and opportunities workshop in conjunction with the 6th Language & Technology Conference;* 56-66

VERGEZ-COURET M., URIELI A. (2014). POS-tagging different varieties of Occitan with single-dialect ressources. Actes de *The First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, Dublin, 21-29.