

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 16^e 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 Oil 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-syntaxique, 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 16^e 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 16^e 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 décrivons 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 16^e siècle, marquée inauguralement par les recueils de textes intitulés *La gente poitevinerie* puis *Le Rolea* au 17^e 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 net¹ 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'empire de Badinghét, qui fâetét 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 11^e 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 phonétiques².

- Pour les premières, on peut notamment mentionner la présence d'un pronom personnel sujet de 1^o et 4^o *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 / dounerae, 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 *fl-* [fj] en graphie normalisée : *pllanjhe* (calme), *bllai* (blé), *cllan* (taon), *glla* (glaçon), *fla* (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 13^e au 20^e 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)	:	<i>ailleurs / aloi / beaucoup/ domaines / fausseté / grand</i>
Gente Poitevinrie 1572	:	<i>ailours / alouay / beacot / demones / faussetez / grons</i>
Gente Poitevinrie 1646	:	<i>ailours / aloiay / beacop / demoynes / fausseti / grond</i>
Dictionnaire P-S Pivetea 1996	:	<i>allour / *aloe / beacop / deménes / faussetez / grand-grant</i>

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 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 PRESTO³. 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'une 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 dictionnaire (cf. figure 1).

Mot n°	Forme rend.	Variante de ...	Lemme Validé	CG Validée	NC	V	NPro	MINEURE	Adj	Inconnu
0	Le	le	le	MINEURE				le(le)		
1	dotour	dotour	doteur	NC	dotour(doto...					
2	medecinou	medecinou	médecin	NC	médecin(m...					
3	qui	qui	qui	MINEURE				qui(qui)		
4	va	alae	alae	V		alae(alae)				
5	vère									INC
6	in	in	in	MINEURE				in(in)		
7	malade	malade	malade	NC	malade(m...					
8	in	in	in	MINEURE				in(in)		
9	gronde	grond	grand	Adj					grand(grond)	
10	necessity	necessity	nécessitai	NC	nécessitai(...					
11	.	.	.	MINEURE				(.)		
12	LE	le	le	MINEURE				le(le)		
13	PEYSAN	peysan	peysan	NC	peysan(pey...					
14	.	.	.	MINEURE				(.)		
15	ve	ve	ve	MINEURE				ve(ve)		
16	me	me	me	MINEURE				me(me)		
17	véré	vère	vère	V		vère(vère)				
18	au	au	au	MINEURE				au(au)		

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 externe⁴ et « voir » comment elle permet de rendre compte des phénomènes observés⁵. 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 VariaLog⁶, 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 16^e 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/disoy*s) 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 mot⁷, (*enfant, renard, renart*), mais pas le reste du temps⁸ ; 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 avons 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 dictionnaire⁹ 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 avons par ailleurs mentionné deux autres phénomènes majeurs, concernant les aspects disons « grammaticaux » du lexique. Il s'agit là de zones « fermées » ou régulières¹⁰ 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^e et 17^e 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'accroître 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 fäscht 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-saintongaise*, n° 8, spé. Langue poitevine-saintongaise, UPCP.
- GAUTHIER P., JAGUENEAU L. (coord.) (2002). Écrire et parler poitevin-saintongeais du XVI^e siècle à nos jours. *Actes du Colloque tenu à l'Université de Poitiers les 26-27 octobre 2001*, Parlanjhe Vivant-Geste éditions.
- GAUTHIER M. (1993). Grammaire du poitevin-saintongeais, Mougou : 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., DUCHET JL. (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, pvtv-stg / français et français / pvtv-stg, Mougou : 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 Castilian 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