Constitution et annotation d'un corpus écrit de contes et récits en occitan

Construction and Annotation of an Occitan Written Narrative Stories Corpus

Marianne Vergez-Couret[1]

[1] Queen's University, Belfast

RÉSUMÉ. Dans cet article, nous présentons les différentes étapes de la constitution d'un corpus de contes en occitan, de l'image numérisée au corpus annoté morphosyntaxiquement. Nous nous focalisons sur deux outils pour l'océrisation du corpus d'une part et pour l'analyse morphosyntaxique d'autre part en mettant en évidence les meilleurs aménagements possibles de ces outils pour la constitution d'un corpus en occitan.

ABSTRACT. In this article, we present the different steps to constitute an Occitan written narrative stories corpus, from digitized images to annotated corpus. We focus on two tools for optical character recognition on one hand and part-of-speech tagging on the other hand, in identifying the best arrangement of those tools to constitute a corpus in Occitan.

MOTS-CLÉS. Corpus, TAL, OCR, Analyse morphosyntaxique, Occitan, Narration, Contes.

KEYWORDS. Corpus, NLP, OCR, Part-Of-Speech Tagging, Occitan, Narrative Stories.

1. Introduction

Dans cet article, nous présentons et discutons notre métholodolgie pour la constitution et l'annotation d'un corpus écrit de contes et récits en occitan, issus de la tradition populaire, publiés du milieu du XIX^{ème} siècle jusqu'au début du XX^{ème} siècle, réalisé au sein du projet ExpressioNarration, subventionné par une Marie Curie Individual Fellowhip (Horizon 2020). Pour constituer ce corpus, que nous appelons OWT (*Occitan, Written, Traditional*), nous nous appuyons sur toutes les avancées méthodologiques en linguistique de corpus en ce qui concerne les outils de numérisation et d'océrisation, d'annotation et de diffusion (format XML TEI P5). Néanmoins, ces outils nécessitent des aménagements que nous discuterons pour tenir compte de la variation graphique et dialectale présente dans notre corpus.

La constitution de ce corpus s'inscrit dans un projet plus large, ExpressioNarration, qui vise à explorer la relation entre complexité linguistique et oralité dans différentes expressions de narration orale en occitan et en français à travers une étude de la temporalité dans un cadre qui rend compte des rapports complexes et variables entre sources orales/écrites et diffusion orale/écrite. Le corpus OWT est constitué de contes et récits issus de collectes de la littérature orale et publiés du milieu du XIX^{ème} siècle jusqu'au début du XX^{ème} siècle. La mise en écrit des contes et récits, transmis de manière orale dans une tradition orale amorce une rupture avec l'oralité [BRU 10]. Contes transmis de manière orale et contes publiés appartiennent à deux systèmes régis par leurs propres régles, l'oral et l'écrit. Les théories linguistiques actuelles ont réussi à dépasser la dichotomie oral/écrit [KOC 01 ; CRY 01]. Et, Carruthers a spécifiquement exploré la façon dont les différents types d'oralité opérent dans le cas particulier de la narration orale à partir d'une analyse d'un corpus de contes et récits en français [CAR 05]. Mais la description des traits des genres oraux peut encore être approfondie, notamment dans le cas de langues minorisées comme l'occitan. Le corpus de contes et récits publiés témoigne de l'entreprise de mise en écrit de données orales. Ce corpus écrit sera comparé à deux autres corpus oraux : un corpus de contes traditionnels (contes transmis par voie orale dans une tradition orale) et un corpus de contes contemporains (contes performés oralement par des conteurs qui puisent leurs sources dans l'écrit et dans l'oral). Nous souhaitons ainsi apporter une contribution originale au débat sur les rapports

oral/écrit en enrichissant les méta-caractéristiques traditionnellement prises en compte dans les modèles existants avec différentes conditions de transmission, de performance et de diffusion. L'accent sera mis sur l'annotation et l'analyse des traits temporels, signalés dans la littérature comme importants pour la structure linguistique des narrations, parmi lesquels les temps verbaux, les connecteurs temporels et les adverbes de localisation temporelle.

Dans cet article, nous nous focalisons sur les premières étapes de la constitution du corpus OWT, de l'image numérisée au corpus pré-annoté morphosyntaxiquement, corpus annoté qui servira de base à l'annotation des traits temporels à l'étude dans le projet. Nous présentons essentiellement deux outils pour l'océrisation du corpus d'une part et pour l'analyse morphosyntaxique d'autre part en mettant en évidence les meilleurs aménagements possibles de ces outils pour la constitution d'un corpus en occitan. Nous discuterons dans la section 2 du choix des sources et des spécificités d'un corpus en langue régionale : l'absence totale de données numériques ; la multitude des graphies employées dans les versions publiées et la variation dialectale. La section 3 sera consacrée à la question de la numérisation et de l'océrisation et la section 4 à l'annotation morphosyntaxique.

2. La littérature orale publiée en occitan

Plusieurs entreprises d'édition ont eu pour objectif la mise en écrit et la publication de contes et récits « ancestraux » [BRU 10]. La mise en écrit entraîne nécessairement des modifications importantes du fait du changement de nature du canal, oral puis écrit. Notamment, les auteur(e)s qui se sont consacré(e)s à cette mise en écrit ont obligatoirement eu recours à divers procédés stylistiques afin de rendre compte des variations de voix, rythme, intensité, gestes et mimiques caractéristiques de l'oral. Chaque conte et récit publié va être ainsi inévitablement plus ou moins fortement marqué du style de l'auteur qui en a fait la transcription. Pour la constitution de ce corpus, notre critère de sélection est la source du conte ou récit publié que nous souhaitons orale en premier lieu. Sont ainsi exclus contes et récits que nous qualifions de « création littéraire », autrement dit qui non pas été orale en premier lieu, même si ces derniers ont été fortement inspirés de la tradition orale. Néanmoins, il est important de noter qu'il n'existe pas de frontière nette entre ces deux sous-ensembles et qu'il est plus prudent de considérer un continuum allant de contes issus de la tradition orale aux créations littéraires. Nous focalisons alors notre attention à l'extrémité du continuum où l'influence de la matière orale lors de la mise en écrit est la plus respectée.

Les ouvrages que nous retenons ont été publiés en occitan (ou en bilingue français/occitan). L'occitan est une langue romane parlée dans le sud de la France, le val d'Aran en Espagne et douze vallées italiennes. Elle se caractérise par une variation interne relativement importante organisée en six dialectes principaux : auvergnat, gascon, languedocien, limousin, provençal et vivaro-alpin. L'occitan n'est pas une langue unifiée et standardisée, et chaque dialecte connait une variation interne propre.

Figure 1.1. *Aire linguistique occitane (Carles, 2005)*

Nous avons pu réunir un corpus de contes et récits issus de plusieurs ouvrages de deux grands ensembles dialectaux, le languedocien et le gascon, mais ce corpus pourra à l'avenir être étendu aux autres ensembles dialectaux pour lesquels des ouvrages de collectes de contes et récits ont également été publiés.

Le corpus a les caractéristiques de l'occitan tel qu'il s'écrit à cette époque-là : à savoir, en plus de la variation dialectale, l'absence de graphie unifiée. Le recul des publications en occitan à la fin du Moyen-Âge entraîne la perte des usages orthographiques médiévaux alors stables. Le milieu du XIX^{ème} siècle voit naître la renaissance de la littérature en occitan autour du Félibrige, fondé par Frédéric Mistral. Ce dernier codifie pour le provençal une orthographe dite "mistralienne" mais cette époque est surtout caractérisée par une multitude de graphies non standards individuelles qui néanmoins partage la caractéristique d'être basée sur les relations phonie/graphie de l'orthographe française, cf. Exemple 1a. Depuis le milieu du XX^{ème} siècle, une nouvelle convention graphique est élaborée inspirée de la graphie des troubadours, elle est dite "graphie classique". Cette graphie a été élaborée avec pour objectif l'atténuation des différences dialectales tout en respectant les particularités de chaque dialecte [SIB 07]. Nous avons opté pour cette graphie pour notre corpus final, cf. Exemple 1b.

Exemple 1. *a) Y aueuo un cop un rey qu'aueuo dus maynatges. (graphie non standard, Bladé)*
b) I avèva un còp un rei qu'avèva dus mainatges. (graphie classique)
c) Il était une fois un roi qui avait deux enfants. (traduction française)

Au début du projet, il n'existe, à notre connaissance, aucune version numérique de ces ouvrages de collectes de contes de tradition orale. Néanmoins, certains d'entre eux, libres de droit, sont disponibles sur les sites des bibliothèques nationales comme Gallica et son équivalent occitan, Occitanica, qui se sont lancées, il y a quelques années, dans des projets de numérisation de grande envergure. Il s'agit d'ouvrages au format pdf constitués d'une image par page d'excellente qualité. Cette image est généralement accompagnée de la sortie d'un logiciel de reconnaissance optique de caractères (OCR, *optical character recognition*) qui n'est pas disponible à la visualisation mais qui permet néanmoins à l'utilisateur de faire des recherches dans le texte. La figure 1.2. est une image extraite de "Contes et proverbes recueillis en Armagnac" par J.-F. Bladé disponible sur le site Occitanica[1] accompagnée de la sortie OCR (logiciel commercial ABBYY FineReader) qui nous a été fournie par le CIRDOC (Centre Interrégional de Développement de l'Occitan), responsable du projet de numérisation Occitanica. La sortie OCR, quoique de très bonne qualité, n'est pas une réplique fidèle du texte de l'image. Pour éditer une version numérique, il faut au préalable corriger les erreurs OCR, rétablir la ponctuation, la mise en forme (cf. les premières lignes de l'image), les paragraphes, les mots tronqués...

3. Océrisation de OWT

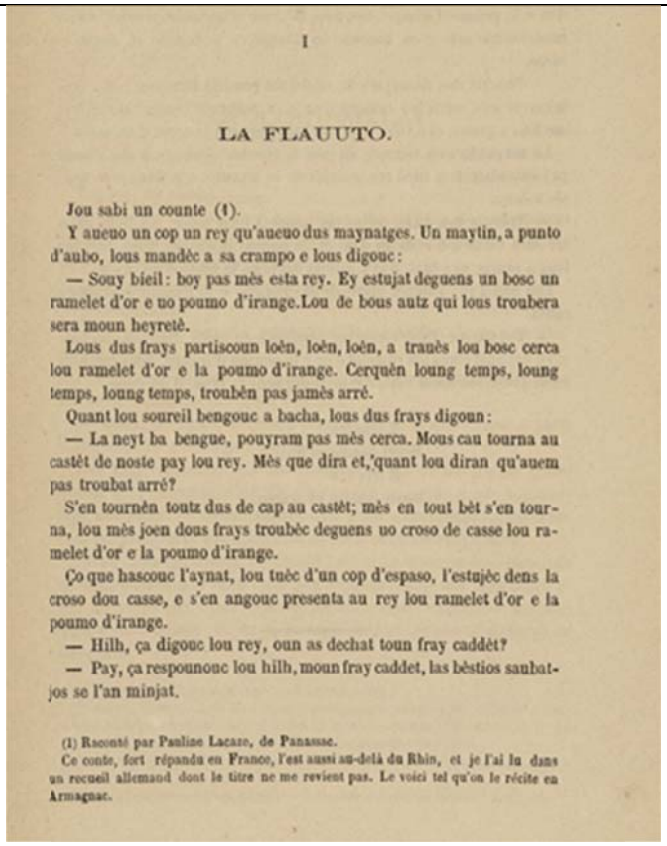
Les ouvrages que nous avons retenus pour notre corpus ont été numérisés par les projets Gallica ou Occitanica avec une très bonne qualité d'image. Dans cette section, nous discuterons uniquement de la question de l'océrisation, i.e. le passage d'une image à du texte brut. La difficulté à surmonter pour l'océrisation d'une langue comme l'occitan est l'absence d'outils adaptés à la langue dans sa variété dialectale et graphique. Nous commençons par recenser les types d'outils possibles pour l'océrisation de l'occitan en détaillant les stratégies employées :

– dans le cadre du projet Occitanica, le CIRDOC sous-traite une société de services qui emploie le logiciel commercial ABBYY FineReader. ABBY FineReader propose un logiciel OCR pour l'occitan, qui n'a pas recours à un dictionnaire, sans préciser ni le dialecte, ni la graphie ;

– d'autres stratégies pour l'océrisation des langues "peu dotées" consistent à utiliser des logiciels tel que OneNote (logiciel commercial développé par Microsoft) ou Tesseract (logiciel libre développé par Google) pour des langues proches graphiquement, i.e. qui disposent du même alphabet [VER 15a] ;

[1] www.occitanica.eu

– enfin, il est possible de recourir à des outils génériques probabilistes, qui utilisent des techniques d'apprentissage automatique supervisé comme Jochre (développé par Joliciel) ou Tesseract. Ces techniques nécessitent la constitution de corpus annotés et optionnellement un lexique [URI 13a, VER 15a].



I

LA FLAUUTO.

Jou sabi un counte (1).
Y aueuo un cop un rey qu'aueuo dus maynatges. Un maytin, a punto d'aubo, lous mandèc a sa crampo e lous digouc:
— Souy bieil: boy pas mès esta rey. Ey estujat deguens un bosc un ramelet d'or e uo poumo d'irange.Lou de bous autz qui lous troubera sera moun heyretè.
Lous dus frays partiscoun loèn, loèn, loèn, a trauès lou bosc cerca lou ramelet d'or e la poumo d'irange. Cerquèn loung temps, loung temps, loung temps, troubèn pas jamès arré.
Quant lou soureil bengouc a bacha, lous dus frays digoun:
— La neyt ba bengue, pouyram pas mès cerca. Mous cau tourna au castèt de noste pay lou rey. Mès que dira et,'quant lou diran qu'auem pas troubat arré?
S'en tournèn toutz dus de cap au castèt; mès en tout bèt s'en tourna, lou mès joen dous frays troubèc deguens uo croso de casse lou ramelet d'or e la poumo d'irange.
Ço que hascouc l'aynat, lou tuèc d'un cop d'espaso, l'estujèc dens la croso dou casse, e s'en angouc presenta au rey lou ramelet d'or e la poumo d'irange.
— Hilh, ça digouc lou rey, oun as dechat toun fray caddèt?
— Pay, ça respounouc lou hilh, moun fray caddet, las bèstios saubatjos se l'an minjat.

(1) Raconté par Pauline Lacaze, de Panassac.
Ce conte, fort répandu en France, l'est aussi au-delà du Rhin, et je l'ai lu dans un recueil allemand dont le titre ne me revient pas. Le voici tel qu'on le récite en Armagnac.

LA FLAUUTO.

Jou sabi un counte (1). Y aueuo un cop un rey qu'aueuo dus maynatges. Un maytin, a punto d'aubo, lous mandèc a sa crampo e lous digouc:

— Souy bieil: boy pas mès esta rey. Ey estujat deguens un bosc un ramelet d'or e uo poumo d'irange.Lou de bous autz qui lous troubera sera moun heyretè.
Lous dus frays partiscoun loèn, loèn, loèn, a trauès lou bosc cerca lou ramelet d'or e la poumo d'irange. Cerquèn loung temps, loung temps, loung temps, troubèn pas jamès arré.
Quant lou soureil bengouc a bacha, lous dus frays digoun:

— La neyt ba bengue, pouyram pas mès cerca. Mous cau tourna au castèt de noste pay lou rey. Mès que dira et/quant lou diran qu'auem pas troubat arré?
S'en tournèn toutz dus de cap au castèt; mès en tout bèt s'en tourna, lou mès joen dous frays troubèc deguens uo croso de casse lou ramelet d'or e la poumo d'irange.
Ço que hascouc l'aynat, lou tuèc d'un cop d'espaso, l'estujèc dens la croso dou casse, e s'en angouc presenta au rey lou ramelet d'or e la poumo d'irange.

— Hilh, ça digouc lou rey, oun as dechat toun fray caddèt?

— Pay, ça respounouc lou hilh, moun fray caddet, las bèstios saubatjos se l'an minjat.
(1) Raconté par Pauline Lacaze, de Panassac.
Ce conte, fort répandu en France, l'est aussi au-delà du Rhin, et je l'ai lu dans un recueil alleraand dont le titre ne me revient pas. Le voici tel qu'on le récite en Arniagnac

Figure 1.2. *Une page extraite de "Contes et proverbes recueillis en Armagnac" par J.-F. Bladé – image et sortie OCR*

Dans cet article, nous proposons de comparer les résultats de 4 logiciels OCR : ABBY FineReader dans sa version occitane, OneNote pour tester le transfert technologique depuis le français, Tesseract pour tester le transfert technologique depuis deux langues, le français et le catalan et Jochre pour un entraînement spécifique avec des ouvrages en occitan et en graphies non standards fournis par le CIRDOC.

Les OCR sont évalués sur deux ouvrages de notre corpus OWT, 10 images de "*Contes populaires recueillis en agenais*" (1874) de Bladé et 8 images de "*Contes et proverbes recueillis en Armagnac*" (1867) du même auteur. Néanmoins, le premier ouvrage est édité en languedocien tandis que le second en gascon. Les sorties Jochre des 18 images ont été corrigées manuellement pour former le corpus de référence.

3.1. *Transfert technologique depuis des langues mieux dotées : choix des langues*

Dans le cas des langues « peu dotées », le recours aux outils d'une langue étymologiquement proche est une stratègie communément adoptée. En ce qui concerne l'océrisation [VER 15a], les auteurs utilisent Tesseract pour l'océrisation de l'occitan (graphie classique) avec les modèles français et catalan et de l'alsacien avec les modèles français et allemand. Ils concluent que, dans le cas de l'OCR, la proximité graphique prime sur la proximité étymologique. Nous commençons donc par évaluer la proximité graphique, i.e. le partage du même ensemble de caractères entre l'occitan (graphies non standards) et les alphabets du français et du catalan, en comparant les caractères présents dans notre corpus d'évaluation aux deux alphabets. Les caractères français couvrent 99,83% du corpus et les

caractères catalans 99,97%. Les modèles français et catalan de Tesseract sont donc de très bons candidats pour assurer le transfert technologique vers l'occitan (graphies non standards). Le tableau 1.1. spécifie les caractères du corpus qui sont spécifiquement français ou spécifiquement catalans. Une fusion des deux modèles, possible avec l'outil Tesseract, permet de couvrir la totalité des caractères du corpus occitan.

Caractères spécifiquement français	6	Caractères spécifiquement catalans	34
ê	2	O	34
ô	2		
Ç	1		
ü	1		

Tableau 1.1. *Caractères alphabétiques dans le corpus d'évaluation occitan*

Nous évaluerons donc le modèle français de OneNote (logiciel commercial), les modèles français et catalan ainsi que le modèle multilingue français-catalan de Tesseract.

3.2. *Entraînement de Jochre*

3.2.1. *Principes de fonctionnement*

La deuxième stratégie que nous adoptons et évaluons, consiste à utiliser des outils génériques probabilistes qui utilisent des méthodes d'apprentissage automatique supervisé, comme Jochre ou Tesseract. Dans cet article, nous nous focalisons sur l'apprentissage d'un modèle occitan (graphies non standards) avec Jochre qui propose la constitution de corpus d'apprentissage à partir d'images réelles. Pour en savoir plus sur l'entraînement de Tesseract à partir d'images générées, nous vous proposons de vous référer à [VER 15a].

Jochre (Java Optical CHaracter REcogition) est un logiciel OCR opensource développé par Assaf Urieli (Joliciel) [URI 13a]. L'analyse avec Jochre s'effectue en trois étapes : la segmentation des images, la reconnaissance des lettres et la correction à l'aide d'un lexique.

La segmentation des images en paragraphes, lignes, mots et « formes » utilise des techniques statistiques *ad hoc*. La figure 1.3. illustre le résultat de cette étape sur une portion d'image. Les « formes » cherchent idéalement à correspondre aux caractères mais cette étape produit parfois des formes qui sont des regroupements de caractères ou des formes qui sont des caractères scindés.

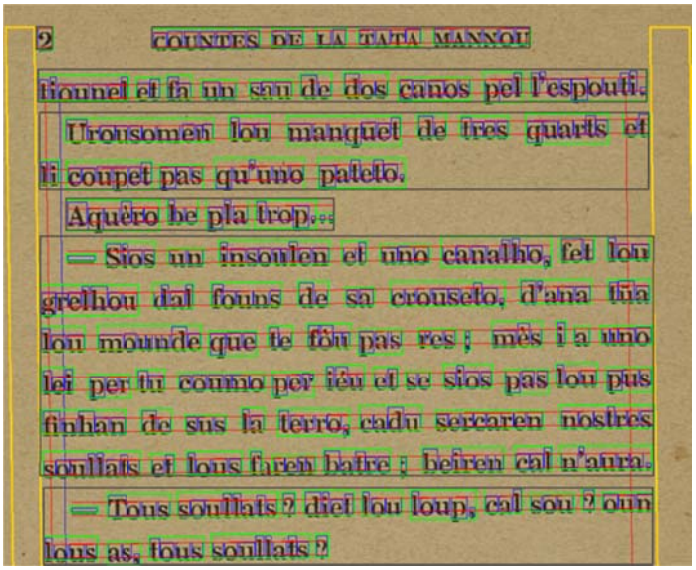


Figure 1.3. *Exemple de segmentation d'un texte en occitan*

La reconnaissance des lettres est le cœur de l'activité de la reconnaissance optique des caractères. Elle s'applique aux formes repérées à l'étape précédente pour leur attribuer la lettre correpondante (plusieurs lettres dans le cas de regroupement de caractères et moitié de la lettre dans le cas des caractères scindés). L'analyse d'une forme dans une nouvelle image utilise un modèle statistique pour lui attribuer une distribution de probabilités de lettres. Ce modèle statistique est construit avec des techniques d'apprentissage supervisé. Cette étape s'appelle l'entraînement ou l'apprentissage. L'utilisateur charge des images scannées (cf. figures 1.2. et 1.3.) dans une application en ligne, JochreWeb et attribue manuellement la bonne lettre à chaque forme (cf. figures 1.4. et 1.5.).

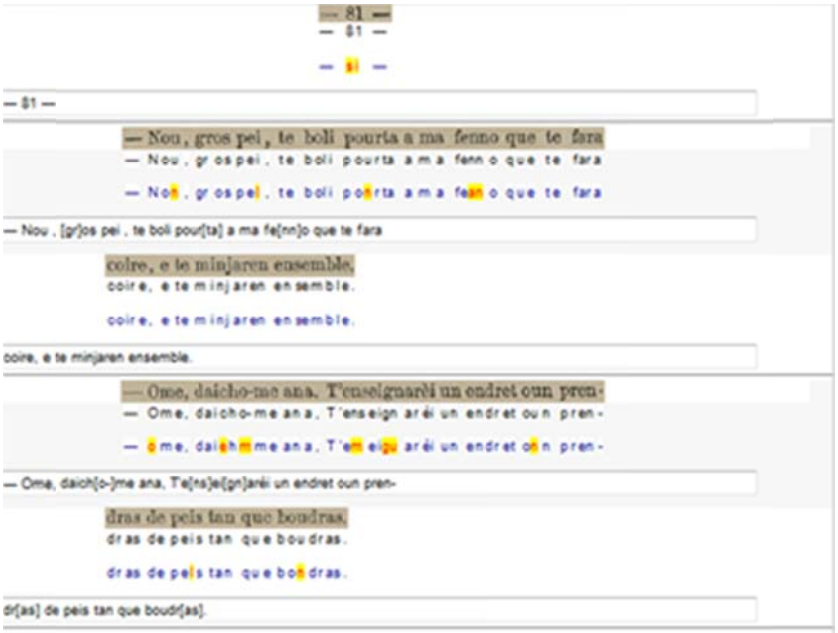


Figure 1.4. *Extrait de l'interface JochreWeb pour la constitution des corpus d'apprentissage*

Figure 1.5. *Attribution manuelle de la (ou les) bonne(s) lettre(s) sur un mot*

Toutes les images annotées en lettres constituent le corpus d'apprentissage sur lequel va être entraîné le modèle statistique. Chaque forme annotée avec une lettre va automatiquement être décrite à l'aide d'une liste de descripteurs (*features* en anglais) : par exemple, la hauteur, la largeur, la présence d'un point… Ensuite, le classifieur robuste (suivant un algorithme itératif complexe, ici un classifieur SVM linéaire [FAN 08]) va mettre en correspondance le résultat des descripteurs avec la ou les lettres favorisées. Par exemple, si la forme est étroite, avec jambage inférieur et un point, la lettre "j" sera probablement favorisée dans le corpus. Le classifieur attribue un poids relatif à chaque résultat de descripteur pour chaque lettre et stocke ces poids dans le modèle statistique, par exemple si la forme contient un point, la probabilité qu'elle soit un "i" est de 95% et la probabilité qu'elle soit un "j" de 5%.

Lors de l'analyse, le classifieur décrit la nouvelle forme à annoter avec les mêmes descripteurs et utilise le modèle statistique pour générer une destribution de probablités pour chaque lettre de l'alphabet et ainsi proposer les *n* analyses les plus probables pour chaque mot.

La dernière étape, dite correction des mots à l'aide du lexique, consiste à reclasser les *n* analyses les plus probables pour chaque mot (aussi appelées faisceau) de l'étape précédente en utilisant un lexique. Un poids plus important sera attribué aux analyses proposant un mot présent dans le lexique. Par exemple, pour l'image  (cf tableau 1.2.), Jochre propose les trois analyses suivantes dans l'ordre de la plus probable à la moins probable : "acordot", "acordat" et "acordet". Le score des mots qui ne sont pas présents dans le lexique ("acordot" et "acordet") est diminué par l'application d'un coefficient de réduction (x 0,5), ce qui entraîne un reclassement, la bonne analyse "acordat" obtenant alors le score le plus élevé.

	Score initial	Présence dans le lexique	Score ajusté
acordot	72,0%	Non (x0,5)	36,0%
acordat	70,1%	Oui (x1,0)	70,1%
acordet	64,3%	Non (x0,5)	32,2%

Tableau 1.2. *Exemple de reclassement de 3 analyses en occitan*

Cette dernière étape est optionnelle. Mais nous choisissons de l'inclure à notre expérience en nous appuyant sur les travaux précédents [URI 13a ; VER 15a] qui ont montré un impact systématiquement positif du lexique sur les résultats de l'OCR quelles que soient la qualité et la couverture de ce dernier.

Nous avons parametré l'évaluation avec un faisceau de 10 analyses les plus probables et un coefficient de réduction de 0,75.

3.2.2. *Préparation des corpus et du lexique*

Le corpus d'entraînement a été constitué, dans le cadre du projet RESTAURE, à partir de 136 pages de 17 œuvres publiées (en moyenne 8 pages par œuvre) du milieu du XIX^{ème} siècle au début du XX^{ème} siècle. Il est varié du point de vue des genres (poésie, prose, préface), de sorte à faire varier la mise en forme, le type, la taille et le style de police. Le corpus est également diversifié du point de vue dialectal, même si de façon inégale, avec 2 textes gascons, 10 textes languedociens et 5 textes provençaux. Le corpus a été annoté manuellement à l'aide de l'outil JochreWeb par trois annotateurs[2].

Le lexique est une liste de formes fléchies et de lemmes constituée à partir :

– des entrées d'un dictionnaire, le *Dictionnaire du béarnais et du gascon modernes* de S. Palay ainsi que des formes fléchies puisées dans les exemples de ce dictionnaire pour le gascon (au total 90 000 lemmes et formes fléchies) ;

– et des formes fléchies issues de 33 œuvres de 17 auteurs de la Base de Textes en langue Occitane, BaTelÒc[3] [BRA 16]. 5 dialectes y sont inégalement représentés (cf. tableau 1.3.) : le languedocien, le limousin, le provençal, le gascon et le vivaro-alpin.

Dialecte	Nombre de formes fléchies
Languedocien	47 242
Gascon	6 246 (BaTelÒc) et 95 232 (BaTelÒc + Palay)
Provençal	50 658
Limousin	5 770
Vivaro-alpin	6 400

Tableau 1.3. *Nombre de formes fléchies par dialecte*

Notre lexique final comprend plus de 190 000 lemmes et formes fléchies. Corpus et lexique sont ici des ressources multidialectales.

3.3. *Evaluation de 4 logiciels OCR*

Dans cette section, nous proposons l'évaluation de 4 logiciels OCR avec pour objectif de choisir celui ou ceux qui nous offre(nt) les meilleurs résultats, autrement dit qui réduit (ou réduisent) au minimum le temps de correction pour la constitution de notre corpus. L'évaluation porte sur les caractères et sur les mots. Elle est effectuée au moyen de deux scripts de comparaison :

– cdiff.py développé par Delphine Bernhard dans le cadre du projet RESTAURE[4] : ce script propose une normalisation de certains caractères tels que les apostrophes, guillemets et tirets et fournit le taux de réussite de reconnaissance des caractères en pourcentage.

– wdiff : programme basé sur *diff* qui permet de comparer et compter le nombre de mots communs dans deux fichiers, le fichier de référence (corrigé manuellement) et la sortie de l'OCR. La mesure donnée dans le tableau 1.2. correspond au pourcentage de mots communs par rapport au nombre total de mots dans le fichier de référence.

[2] Nous remercions les 3 stagiaires, Lucie Bergé, Eunbee Kang et Estelle Pompon, qui ont participé à la constitution des corpus d'annotation dans le cadre de stages RESTAURE (ANR-14-CE24-0003).
[3] http://redac.univ-tlse2.fr/bateloc/index.jsp
[4] ANR-14-CE24-0003

Nous comparons notre corpus de référence, à la sortie d'AbbyFineReader fourni par le CIRDOC (ABBYY), la sortie de OneNote avec le modèle "français" (OneNote), les sorties de Tesseract avec le modèle "français" (TESS-FRA), le modèle "catalan" (TESS-CAT) et le modèle multilingue "français-catalan" (TESS-FRACAT) et enfin la sortie de Jochre après l'entraînement décrit ci-dessus (Jochre).

	Pourcentage d'erreurs au niveau des caractères (après normalisation)	Pourcentage des mots communs (après normalisation)
ABBYY	1,58%	91%
OneNote	13,25%	75%
TESS-FRA	21,91%	68%
TESS-CAT	21,38%	69%
TESS-FRACAT	21,87%	70%
Jochre	1,42%	94%

Tableau 1.4. *Résultats en pourcentage des performances des logiciels OCR*

Les résultats montrent que les modèles proposés par Tesseract et OneNote obtiennent des résultats bien en deça des autres logiciels OCR avec des taux d'erreur sur le caractère dépassant les 10%, voire les 20% dans le cas de Tesseract. Dès que des logiciels sont construits spécifiquement pour la langue à océriser, il est tout à fait normal d'espérer de meilleurs résultats, que nous obtenons avec ABBYY et Jochre. Les deux logiciels obtiennent des résultats très convenables avec un taux d'erreur inférieur à 2% pour les caractères et à 10% pour les mots. Le temps consacré à la constitution des corpus et du lexique pour l'océrisation de l'occitan (graphies non standards) avec Jochre est un bon investissement au vue de l'amélioration des résultats que cela permet par rapport aux stratégies de transfert technologique depuis une langue mieux dotée et pour doter la langue d'un outil opensource.

Pour cette expérience, seules les graphies non standards basées sur les rapports phonies/graphies du français ont été prises en compte. Néanmoins, toutes les variations graphiques individuelles et dialectales ont été ignorées dans la gestion des ressources. La prise en compte de ces deux types de variation est présentée dans la section suivante avec le logiciel Jochre et les ressources qui ont été constituées, cf. section 3.2.2.

3.4. *Evaluation par dialecte avec Jochre*

Nous proposons l'évaluation de plusieurs modèles entraînés avec Jochre faisant varier l'utilisation des corpus annotés et des lexiques unidialectal vs. multidialectal. Les modèles sont les suivants :

	Commentaires	Pages pour l'entraînement
P34L83G18	Modèle de base	Provençal : 34 pages (5 œuvres différentes) Languedocien : 83 pages (10 œuvres différentes) Gascon : 18 pages (2 œuvres différentes)
L83 (modèle uniquement avec des textes languedociens)	Modèle languedocien Variété des œuvres	Languedocien : 83 pages (10 œuvres différentes)
G83 (modèle uniquement avec deux textes gascons)	Modèle gascon Absence de variété	Gascon : 83 pages (2 œuvres différentes)
L83G83	Modèle combiné languedocien et gascon Surreprésentation des deux œuvres en gascon	Languedocien : 83 pages (10 œuvres différentes) Gascon : 83 pages (2 œuvres différentes)
P34L83G83	Modèle combiné multidialectal (languedocien, gascon et provençal)	Provençal : 34 pages (5 œuvres différentes) Languedocien : 83 pages (10 œuvres différentes) Gascon : 83 pages (2 œuvres différentes)
P34L83G175	Modèle combiné multidialectal Très grande surreprésentation des deux œuvres en gascon	Provençal : 34 pages (5 œuvres différentes) Languedocien : 83 pages (10 œuvres différentes) Gascon : 175 pages (2 œuvres différentes)

Tableau 1.5. *Modèles entraînés avec Jochre*

Le modèle P34L83G18 est utilisé pour l'évaluation précédente, cf. section 3.2. et 3.3. Le modèle L83 est construit uniquement avec des textes languedociens de 10 œuvres différentes, garantissant une bonne variété. Le modèle G83 est construit avec uniquement 2 œuvres en gascon. Il est constitué du même nombre de pages que le modèle précédent mais avec une absence de variété. Le modèle L83G83 est un modèle combiné avec le même nombre de pages en languedocien et en gascon mais avec une variété d'œuvres plus importante pour le languedocien que pour le gascon. Le modèle P34L83G83 est un modèle multidialectal, avec des œuvres en provençal en plus des œuvres du modèle précédent. Enfin le modèle P34L83G175 utilise toutes les pages disponibles pour l'entraînement. Les deux seules œuvres en gascon sont alors surreprésentées. Nous cherchons à répondre aux questions suivantes :

– Est-ce qu'un ciblage du dialecte du texte à océriser permet une amélioration des résultats (cf. modèles L83 et G83) ?

– Est-ce que la variété des œuvres disponibles est un facteur sur la qualité des résultats (cf. modèles L83 et G83) ?

– Faut-il favoriser une répartition équitable du nombre de pages par œuvre (cf. modèles P34L83G18 vs. P34L83G83 et P34L83G175) ?

L'évaluation porte sur les mots. Elle est effectuée au moyen du programme wdiff (cf. section 3.3.). Nous avons évalué tous les modèles avec le lexique intégral. Pour les modèles L83, G83 et L83G83, nous avons également évalué avec les lexiques du dialecte, respectivement uniquement languedocien pour L83, uniquement gascon pour G83 et combinaison des lexiques languedocien et gascon pour L83G83. En l'absence d'effet sur les résultats, nous ne présentons que les résultats des évaluations faites avec le lexique intégral.

Modèle	Agenais (languedocien) (Pourcentage des mots communs)	Armagnac (gascon) (Pourcentage des mots communs)	Total (Pourcentage des mots communs)
P34L83G18	93,5%	97,2%	95,0%
L83	93,5%	96,9%	94,9%
G83	67,3%	76,8%	71,2%
L83G83	93,1%	96,7%	94,5%
P34L83G83	93,3%	96,9%	94,8%
P34L83G175	93,1%	96,8%	94,6%

Tableau 1.6. *Résultats (en pourcentage de mots communs)*

Le modèle permettant d'obtenir les meilleurs résultats avec Jochre est le modèle de base (utilisé pour l'évaluation précédente) avec une variété d'œuvres et une répartition équitable des pages par œuvre. La baisse importante des résultats (de 95% à 71%) avec le modèle ne répondant pas à ces deux critères (G83) est très significative. En revanche, tous les autres résultats ne présentent pas de différences significatives. Nous en concluons qu'il n'y a pas d'effets significatifs du ciblage du dialecte sur les performances de Jochre avec les ressources dont nous disposons actuellement.

3.5. *Correction des sorties OCR*

Les sorties des deux logiciels ABBYY et Jochre ont été exploitées pour constituer nos corpus de sorte à tester deux procédures de correction. Les sorties ABBYY, qui sont des textes bruts ont été corrigées avec un éditeur de texte tandis que les sorties de Jochre ont été corrigées via l'inteface JochreWeb (cf. figure 1.4.). A l'issu de ces corrections, la correctrice[5] a déclaré trouver la correction plus performante et plus confortable *via* l'interface JochreWeb. Pourtant, il faut noter que cette interface a été conçue, non pas pour corriger des œuvres entières, mais pour constituer les corpus d'entraînement. Les corpus d'entraînement ont de fait été augmentés lors de cette phase de correction même si cela n'a pas permis d'améliorer les résultats (cf. section 3.4.). Il faudrait approfondir la question des procédures de correction, notamment en mesurant le temps nécessaire et le taux d'erreur restant en employant ces deux procédures par plusieurs correcteurs. Plusieurs œuvres complètes ont été intégralement corrigées et encodées au format xml. Ces textes seront consultables *via* BaTelÒc [BRA 16]. Pour le projet ExpressioNarration, nous avons sélectionné dans ces œuvres une vingtaine de contes et récits que nous avons conservée, dans un premier temps, au format texte brut.

L'étape suivante de la préparation du corpus a consisté en une harmonisation des textes avec l'adoption de la graphie classique qui permet de neutraliser la variation graphique individuelle. La graphie classique est également adoptée pour les transcriptions des deux sous-corpus oraux, garantissant une harmonisation globale du point de vue de la graphie pour les trois sous-corpus. Le passage de la graphie originale à la graphie classique a été fait manuellement. Il ne s'agit pas d'une normalisation vers des formes fréquentes de chaque dialecte. La transcription cherche autant que possible à respecter les particularités du parler de chaque auteur et ne cherche pas à neutraliser la variation dialectale.

Cette harmonisation permet également de faciliter la mise en œuvre de l'étape suivante, à savoir l'analyse morphosyntaxique automatique, pour laquelle des outils et des ressources ont déjà été développés pour l'occitan en graphie classique.

[5] Nous remercions Lucie Bergé qui a effectué ces corrections dans le cadre d'un stage financé par le projet RESTAURE (ANR-14-CE24-0003).

4. Analyse morphosyntaxique du corpus OWT

Cette étape consiste à attribuer automatiquement une étiquette morphosyntaxique à chaque mot du corpus : le lemme, la partie du discours (nom, adjectif, adverbe, verbe…) et des informations morphosyntaxiques associées (genre, nombre, mode, temps…). Cette annotation automatique est, pour la constitution de notre corpus, un préalable à l'annotation des traits temporels (cf. conclusion). Un exemple d'annotation est présenté dans le tableau 1.7. sur la phrase "Qu'i avè un còp un òmi, s'aperèva Loïson, e ua hemna, s'apèreva Mariolic, qu'èran maridats" (*Il était une fois un homme, il s'appellait Louison, et une femme, elle s'appellait Marioulic, ils étaient mariés*).

Forme	Lemme	Catégorie principale	Informations morphosyntaxiques	Forme	Lemme	Catégorie principale	Informations morphosyntaxiques
Qu'	que	R	pp--	ua	un	D	a-fs-i
i	i	P	p3msd-	hemna	hemna	N	Cfs
avè	aver	V	mi-i3s-	,	,	F	
un	un	D	a-ms-i	s'	se	P	x3ms-
còp	còp	N	cms	aperèva	aperar	V	mi-i3s-
un	un	D	a-ms-i	Mariolic	Mariolic	N	Pfs
òmi	òmi	N	cms	,	,	F	
,	,	F		qu'	que	R	pp--
s'	se	P	x3ms-	èran	èster	V	mi-i3p
aperèva	aperar	V	mi-i3s-	maridats	maridar	A	Fpmp
Loïson	Loïson	N	pms	amassa	amassa	R	gp--
,	,	F		.	.	F	
e	e	C	C				

Tableau 1.7. *Exemple d'annotations morphosyntaxiques sur une phrase*

Le jeu d'étiquettes morphosyntaxiques est adapté du jeu d'étiquettes standard GRACE [RAJ 97], lui-même adapté des jeux d'étiquettes MULTEXT [IDE 94] et EAGLES [REC 96]. La description du jeu d'étiquettes complet pour l'occitan est décrit dans [VER 16].

4.1. *Stratégies*

L'analyse morphosyntaxique automatique pour l'occitan en graphie classique s'est considérablement développée dans le cadre des projets BaTelÒc et RESTAURE. De nombreuses ressources, que nous exploiterons pour cette expérience, ont ainsi été consistuées. Elles nous permettent d'adopter des méthodes par apprentissage supervisé qui sont à l'heure actuelle fortement plebiscitées pour le développement des outils de traitement automatique des langues (TAL). Des outils d'analyse morphosyntaxique automatique ont ainsi été développés pour des langues standardisées comme le français, l'anglais ou l'allemand. Néanmoins, nous disposons de peu d'éléments sur la façon d'adapter ces outils à une langue comme l'occitan connaissant une grande variation interne.

Une stratégie consiste à traiter chaque dialecte ou chaque parler comme une langue à part entière et de constituer pour ce dialecte des ressources spécifiques de très grande qualité (corpus annotés, lexiques ou grammaires selon l'approche choisie), ce qui nécessite des moyens humains et financiers très importants. Une autre stratégie consiste à exploiter les similarités de la langue à annoter à une autre langue mieux dotée proche étymologiquement. Plusieurs méthodes ont été décrites dans la littérature,

comme l'utilisation de textes alignés (bitextes) [TAC 13], l'application de règles sur les caractères pour rapprocher graphiquement la langue à annoter à la langue mieux dotée [HAN 11], et l'identification de cognats lexicaux [SCH 13] ou la création de lexiques bilingues de mots grammaticaux qui sont transposés de la langue mieux dotée vers la langue à annoter [BER 13 ; VER 13]. Ces méthodes exploitent ainsi les outils créés pour d'autres langues. Elles permettent de créer des ressources pour les langues peu dotées à moindre coût mais les ressources ainsi créées sont de moins bonne qualité.

En ce qui concerne l'occitan, l'existence de ressources (corpus annotés et lexiques) ont permis la création de modèles spécifiquement entraînés pour l'occitan avec le logiciel Talismane [URI 13b]. Le premier modèle de Talismane spécifiquement entraîné pour l'occitan est décrit dans [VER 14] avec un corpus d'entraînement de 2500 mots pour un dialecte, et plus précisement le parler du Rouergue, et un lexique de 225 000 formes fléchies pour le languedocien. Un deuxième modèle est décrit dans [VER 15b] avec un corpus étendu à plusieurs dialectes du languedocien de 3000 mots et un lexique de 280 000 formes fléchies pour le languedocien. De nouvelles ressources ont été constituées dans le cadre du projet RESTAURE, en suivant la logique suivante : étendre la création de ressources à d'autres dialectes languedociens, avant d'étendre la création de ressources à un autre dialecte, le gascon.

Dans le projet ExpressioNarration, le corpus OWT contient des textes de deux grands ensembles dialectaux, le languedocien et le gascon. Les expériences que nous présentons dans cet article visent à mettre au jour la meilleure stratégie pour l'annotation morphosyntaxique de ces deux dialectes de l'occitan avec Talismane et les ressources disponibles à l'heure actuelle. Notre stratégie vise ainsi à exploiter les similarités des dialectes pour passer d'un dialecte de l'occitan à un autre.

4.2. *Fonctionnement de Talismane*

Talismane est un logiciel opensource dédié à l'analyse syntaxique [URI 13b]. Pour cette expérience, nous utilisons uniquement le module consacré à l'analyse morphosyntaxique. Talismane utilise des méthodes d'apprentissage automatique supervisé, nécessitant des données annotées pour l'entraînement et optionnellement un lexique.

Chaque forme annotée avec une étiquette morphosyntaxique va automatiquement être décrite à l'aide d'une liste de descripteurs : W la forme lexicale exacte, P l'étiquette attribuée à la forme Tj (si $j < i$) ou les étiquettes trouvées dans le lexique (si $j \geq i$), U si la forme est inconnue dans le lexique, L le lemme trouvé dans le lexique le cas échéant, Sfxn les *n* dernières lettres de la forme (*n* de 2 à 5), 1st si la forme est la première de la phrase, Last si la forme est le dernière de la phrase. Ces briques de base sont aussi combinées en bigrammes et trigrammes. Ainsi, par exemple, Pi-1 estime la distribution des probabilités pour l'étiquette de la forme actuelle étant donné l'étiquette attribuée à la forme précédente. Pi-2Pi-1 estime la distribution des probabilités pour cette même étiquette étant donné les étiquettes attribuées aux deux formes précédentes.

Le lexique intégré à Talismane est utilisé à la fois sous forme de descripteurs (cf. les descripteurs Tj, U et L) et sous forme de règles. Dans ce cas, le lexique permet de contourner les choix du modèle statistique pendant l'analyse, soit en interdisant le choix d'une étiquette : si un mot est uniquement listé dans le lexique des classes fermées (préposition, conjonction, pronom, déterminant…), Talismane interdit le choix d'une étiquette de classes ouvertes (nom, verbe, adjectif…) ; soit en imposant le choix d'une étiquette : Talismane impose de ne jamais choisir une étiquette de classes fermées à moins que le mot à analyser ne soit présent dans le lexique des classes fermées. Les résultats de travaux précédents [VER 14] ont montré l'importance des lexiques, et notamment du lexique des classes fermées, tant au niveau de l'entraînement (lexique utilisé sous forme de descripteurs) que de l'analyse (lexique utilisé sous forme de règles). Dans l'expérience décrite ci-dessous, nous utiliserons systématiquement un lexique mais nous opposerons les résultats obtenus avec un lexique unidialectal (languedocien ou gascon) et un lexique multidialectal (languedocien et gascon).

Talismane permet de combiner plusieurs options traditionnellement utilisées en apprentissage supervisé. Nous avons opté pour un classifieur SVM linéaire avec ε = 0,1 et C = 0,5 [FAN 08].

4.3. *Ressources*

Nous allons présenter dans cette section les ressources que nous avons utilisées pour déterminer la ou les meilleures stratégies pour annoter notre corpus OWT en languedocien et en gascon.

4.3.1. *Lexiques et jeu d'étiquettes (tagset)*

Nous avons utilisé deux lexiques, un lexique du languedocien et un lexique du gascon. Les deux lexiques contiennent pour chaque entrée, le lemme et l'étiquette morphosyntaxique (standard EAGLES [RAJ 97]) avec la catégorie principale et des informations morphosyntaxiques. Les catégories principales sont listées dans le tableau 1.8.

Catégorie principale	Tag
Nom	N
Verbe	V
Pronom	P
Adjectif	A
Déterminant	D
Adverbe	R
Préposition	S
Conjonction	C
Interjection	I
Résidu	X
Ponctuation	F

Tableau 1.8. *Tagset EAGLES*

Le lexique du languedocien que nous avons utilisé correspond à la première version de Loflòc (Lexique Ouvert Fléchi Occitan) [VER 16]. Il a été construit à partir de plusieurs ressources disponibles au format numérique : le Dictionnaire Occitan-Français Languedocien de Laux (2001), le Dictionnaire Français-Occitan Languedocien de Laux (2005), les données de l'application verb'Òc, conjugueur édité par Lo Congrès [SAU 95 ; SAU 16] ainsi que les noms propres du lexique Apertium [ARM 06]. Il contient environ 680 000 formes (en tenant compte uniquement de la catégorie principale).

Le lexique du gascon que nous avons utilisé a été construit à partir du Dictionnaire Français/Occitan (gascon) de Per Noste (2007), pour intégrer Loflòc. Il contient environ 760 000 formes (en tenant compte uniquement de la catégorie principale) mais la qualité de ce lexique est moindre à cause de l'absence à ce jour de correction sur les informations grammaticales issues du dictionnaire qui sont parfois incomplètes ou inadaptées.

La réunion des deux lexiques permet d'obtenir un lexique d'environ 1 200 000 formes, indiquant que le nombre de formes communes aux deux lexiques est assez faible, environ 20% du lexique global.

4.3.2. *Corpus d'entraînement*

Les corpus d'entraînement ont été constitués dans le cadre du projet RESTAURE[6]. Ils ont été annotés manuellement avec le lemme, la catégorie grammaticale et les informations morphosyntaxiques (genre, nombre, personne, temps, mode…) selon le standard EAGLES.

Le corpus languedocien est composé de 6 œuvres :

Œuvres	Extrait annoté (en nombre de mots)
Dels Camins Bartassièrs, M. Esquieu, 2003	512
E la Barta Floriguet, E. Mouly, 1948	8749
Los crocants de Roergue, F. Déléris, 1992	2725
Lo balestrièr de Miramont, R. Marty, 2006	6402
Contes del Drac, J. Boudou, 1975	4207
D'al Brès à la Toumbo, J. Bessou, 2004	5173
Total	**27768**

Tableau 1.9. *Composition du corpus annoté en languedocien*

Le corpus gascon est composé de 3 œuvres :

Œuvres	Extrait annoté (en nombre de mots)
Contes de Gasconha Purmèra Garba, J.-F. Bladé, 2010	5077
Hont Blanca, J.-L. Lavit, 2000	5000
De la pèth de Cohet, A. Peyroutet, 2003	4999
Total	**15076**

Tableau 1.10. *Composition du corpus annoté en gascon*

4.3.3. *Corpus d'évaluation*

Le corpus d'évaluation, appelé OWT-tag, a été constitué dans le cadre du projet ExpressioNarration à partir d'extraits de 4 auteurs. Il a également été annoté manuellement avec le lemme, la catégorie grammaticale et les informations morphosyntaxiques selon les standards EAGLES. OWT-tag est composé de 5 œuvres des deux dialectes, languedocien et gascon :

[6] Nous remercions les stagiaires qui ont participé à l'annotation des corpus : Estel Llansana, Sébastien Gonzales et Aurélie Abadie.

Œuvres	Dialecte	Extrait annoté (en nombre de mots)
Contes populaires recueillis en agenais, J.-F. Bladé, 1874	Languedocien	1503
Contes populaires du Languedoc, L. Lambert, 1880	Languedocien	1509
Contes et proverbes recueillis en Armagnac, J.-F. Bladé, 1867	Gascon	1505
Coundes biarnés, couéilhuts aüs parsàas miéytadès dou péys de Biarn, J.-V. Lalanne, 1890	Gascon	1498
Contes populaires recueillis dans la Grande-Lande, F. Arnaudin, 1887	Gascon	1503
Total		**7518**

Tableau 1.11. *Composition du corpus d'évaluation, OWT-tag*

4.4. *Evaluation*

L'évaluation porte uniquement sur la catégorie principale (cf. tableau 1.8.). Une fois cette catégorie déterminée par Talismane, les informations morphosyntaxiques sont généralement disponibles dans le lexique. Toutes les ressources présentées ci-dessus ont été rassemblées pour répondre aux questions suivantes :

– Peut-on obtenir des résultats satisfaisants avec les ressources actuellement disponibles pour le languedocien et pour le gascon ?

– Peut-on améliorer la qualité des résultats en combinant les lexiques des deux dialectes ?

– Peut-on améliorer la qualité des résultats en combinant les corpus des deux dialectes ?

– Demeure-t-il des variations dans la qualité des résultats selon les dialectes et selon les œuvres ?

Pour répondre à ces questions, 7 modèles ont été entraînés avec Talismane avec différents accès aux ressources, cf. tableau 1.12. Néanmoins dans les sections suivantes, nous ne présentons les résultats que pour les modèles les plus pertinents.

Modèles	Corpus	Lexiques
Corpus-all_Lex-all	Languedocien/Gascon	Languedocien/Gascon
Corpus-lan_Lex-lan	Languedocien	Languedocien
Corpus-gas_Lex-gas	Gascon	Gascon
Corpus-all_Lex-lan	Languedocien/Gascon	Languedocien
Corpus-all_Lex-gas	Languedocien/Gascon	Gascon
Corpus-lan_Lex-all	Languedocien	Languedocien/Gascon
Corpus-gas_Lex-all	Gascon	Languedocien/Gascon

Tableau 1.12. *Modèles Talismane*

4.4.1. *Evaluation globale pour le languedocien*

Dans cette section, nous présentons les résultats obtenus pour les deux textes languedociens réunis. Nous ordonnons ces résultats en respectant l'ordre chronologique de disponibilité des ressources (cf. Section 4.1. sur les stratégies de constitution des ressources).

	Corpus-lan_Lex-lan	Corpus-all_Lex-lan	Corpus-lan_Lex-all	Corpus-all_Lex-all
Résultats en %	92,87	**93,12**	92,49	92,98

Tableau 1.13. *Résultats globaux pour le sous-corpus languedocien*

Au tout départ les premières ressources réunies pour l'annotation du languedocien étaient un corpus et Loflòc, qui sont des ressources spécifiques pour le languedocien (« Corpus-lan_Lex-lan »). Ces ressources permettent d'obtenir un score de 92,87 % d'étiquettes correctes. Puis, des ressources ont été créées pour le gascon, d'abord un corpus que nous utilisons, en plus des autres ressources, dans le modèle « Corpus-all_Lex-Lan », ce qui permet une hausse sensible des résultats avec un score de 93,12%, puis un lexique que nous utilisons dans le modèle « Corpus-lan_Lex-all », ce qui ne permet pas d'amélioration (92,49%). Enfin, nous évaluons un modèle utilisant toutes les ressources disponibles (« Corpus-all_Lex-all »), ce qui nous permet d'obtenir le résultat de 92,98%. Pour conclure, nous obtenons des résultats satisfaisants (> 92%) pour le languedocien avec les ressources du dialecte. Ce score peut légèrement être amélioré grâce à l'utilisation d'un corpus multidialectal qui permet de dépasser les 93%. En revanche, il est préférable d'avoir recours à un lexique unidialectal.

4.4.2. *Evaluation globale pour le gascon*

Dans cette section, nous présentons les résultats obtenus pour les trois textes gascons. De la même façon que précédemment, nous ordonnons ces résultats en respectant l'ordre chronologique de constitution des ressources et en partant des ressources disponibles pour le languedocien.

	Corpus-lan Lex-lan	Corpus-all Lex-lan	Corpus-all Lex-all	Corpus-gas Lex-gas	Corpus-all Lex-gas	Corpus-gas Lex-all
Résultats en %	79,05	82,48	**88,96**	87,82	88,15	87,38

Tableau 1.14. *Résultats globaux pour le sous-corpus gascon*

Sans surprise, d'un modèle qui n'utilise que des ressources en languedocien (« Corpus-lan_Lex-lan ») à un modèle qui utilise en plus un corpus d'entraînement gascon (« Corpus-all_Lex-all »), nous observons une hausse des résultats de 79,05% à 82,48%. Ce score est encore amélioré (87,82%) avec l'utilisation supplémentaire du lexique gascon (« Corpus-all_Lex-all »). Aucune autre combinaison ne permet d'améliorer ce score. Le meilleur résultat pour le gascon (> 88%) est très en deça des résultats obtenus pour le languedocien, en raison probablement de la taille plus petite du corpus d'entraînement garscon et la qualité moindre du lexique gascon.

Enfin, les trois derniers modèles du tableau permettent de montrer un phénomène similaire à celui observé pour le languedocien. En partant d'un modèle unidialectal (« Corpus-gas_Lex-gas ») avec lequel nous obtenons un score de 87,82%, une amélioration est possible avec l'utilisation d'un corpus multidialectal, cf. le score de 88,15% avec le modèle « Corpus-all_Lex-gas » mais pas avec un lexique multidialectal, cf. le score de 87,38% avec le modèle « Corpus-gas_Lex-all ». Nous pouvons établir à partir de cela une stratégie de constitution des ressources : annoter du lexique permet d'améliorer les scores pour le dialecte du lexique tandis qu'annoter un corpus unidialectal permet d'améliorer les scores pour tous les dialectes confondus.

4.4.4. *Evaluation par œuvre*

Dans cette section, nous souhaitons compléter les résultats précédents par une évaluation des modèles par œuvre pour savoir si de meilleures stratégies peuvent être établies en fonction du parler de chaque auteur. Nous commençons par présenter les résultats pour les deux textes languedociens :

	Corpus-lan_Lex-lan	Corpus-lan_Lex-all	Corpus-all_Lex-lan	Corpus-all_Lex-all
Bladé_Agenais	91,81	91,95	92,09	**92,66**
Lambert_Languedoc	93,94	93,04	**94,15**	93,31

Tableau 1.15. *Résultats en % par texte en languedocien*

Les écarts de performance entre les deux textes languedociens tendent à indiquer que les ressources jusque-là constituées sont plus en adéquation avec le parler du territoire du languedoc central qu'avec celui du territoire agenais. Effectivement, la majorité des auteurs des dictionnaires et œuvres utilisés pour constituer les ressources sont originaires de ce territoire (C. Laux (Tarn) et R. Marty (Haute-Garonne)) ou de l'Aveyron tout proche (F. Déléris, J. Boudou, J. Bessou, E. Mouly).

Cela a également une incidence directe sur la meilleure stratégie à adopter pour ces deux textes. Le meilleur score pour Lambert_Languedoc est obtenu avec le modèle qui fait uniquement appel au lexique unidialectal en languedocien (« Corpus-all_Lex-lan ») tandis que le meilleur score pour Bladé_Agenais est obtenu avec le modèle qui fait appel à toutes les ressources disponibles (« Corpus-all_Lex-all »).

Nous procédons de même avec les textes en gascon :

	Corpus-gas_Lex-gas	Corpus-gas_Lex-all	Corpus-all_Lex-gas	Corpus-all_Lex-all
Bladé_Armagnac	85,37	85,51	87,90	**89,97**
Lalanne_Béarn	**89,94**	89,24	89,66	89,03
Arnaudin_Lande	**88,17**	87,39	86,9	87,89

Tableau 1.16 *Résultats en % par texte en gascon*

La meilleure stratégie à adopter pour l'annotation en gascon varie également selon les textes. Pour l'un d'entre eux, Bladé_Armagnac, le meilleur score est obtenu en mobilisant toutes les ressources disponibles comme pour Bladé_Agenais. Les deux ouvrages sont du même auteur, bien que l'un soit en gascon et l'autre en languedocien. Néanmoins, ils proviennent de territoires situés à la frontière linguistique entre les deux dialectes. Il n'est donc pas surprenant que ces deux textes bénéficient des ressources multidialectales puisqu'ils proviennent de zones de transition entre les deux dialectes. Pour les deux autres textes, Lalanne_Béarn et Arnaudin_Lande, qui proviennent de territoires éloignés de la frontière linguistique, les meilleurs scores sont obtenus en mobilisant uniquement les ressources du gascon, tant pour le corpus que pour le lexique.

Ces expériences nous permettent d'établir la meilleure stratégie pour l'annotation automatique de deux dialectes avec les ressources disponibles à ce jour. Nous avons montré qu'il est possible d'adopter une démarche globale (en employant toutes les ressources disponibles) qui permet d'obtenir un résultat supérieur à 92% pour le languedocien et 88% pour le gascon. Ces résultats peuvent être sensiblement améliorés à l'échelle d'une œuvre en tenant compte 1) de la proximité dialectale de l'oeuvre par rapport aux ressources utilisées pour l'entraînement et 2) de la provenance géographique de l'œuvre plus ou moins éloignée d'une frontière dialectale. Elles nous permettent également de choisir le modèle garantissant la meilleure annotation possible pour chaque texte de notre corpus OWT.

5. Conclusion

Dans cet article, nous avons présenté les différentes étapes de la constitution d'un corpus de contes en occitan, de l'image numérisée au corpus annoté. Nous avons établi pour deux outils de traitement automatique des langues, un logiciel de reconnaissance optique de caractères et un analyseur morphosyntaxique, les meilleures stratégies avec les ressources disponibles actuellement pour la constitution de corpus en occitan, en graphies non standards pour l'OCR et en graphie classique pour l'analyseur morphosyntaxique.

Concernant le logiciel de reconnaissance optique de caractères, nos résultats confirment que la variation dialectale peut être ignorée étant donné que la proximité graphique prime (i.e. tous les dialectes s'écrivent avec le même alphabet). Nous avons construit un modèle d'océrisation de textes en occitan graphies non standards avec le logiciel Jochre qui atteint des résultats très convenables avec un taux d'erreur inférieur à 2% pour les caractères et à 10% pour les mots. A l'heure actuelle, la meilleure stratégie pour la constitution d'un modèle OCR consiste à varier la provenance des images, en moyenne 8 pages par œuvre.

Concernant l'analyseur morphosyntaxique, nous avons construit un modèle en occitan graphie classique avec le logiciel Talismane qui permet d'obtenir un résultat supérieur à 92% pour le languedocien et 88% pour le gascon. De plus, nous avons montré que la prise en compte de la proximité dialectale de l'oeuvre par rapport aux ressources utilisées pour l'entraînement et de la provenance géographique de l'œuvre plus ou moins éloignée d'une frontière dialectale sont des facteurs permettant une hausse sensible des résultats. Enfin, nous avons également mis au jour une stratégie de constitution des ressources : la constitution d'un lexique unidialectal permet la hausse des performances pour le dialecte tandis que la constitution d'un corpus annoté unidialectal pour l'entraînement permet la hausse des performances pour tous les dialectes. Ces résultats, bien qu'encourageants, sont bien en deça des résultats obtenus à ce jour pour des langues mieux dotées, comme le français ou l'anglais, pour lesquelles les performances atteintes sont aux alentours de 97%. Une analyse fine des erreurs pourrait permettre de mieux cibler les efforts à fournir pour la constitution de nouvelles ressources.

La prochaine étape de la constitution de ce corpus concerne l'annotation des traits temporels. Elle sera effectuée manuellement mais en s'appuyant sur une pré-annotation automatique produite à partir des sorties de Talismane pour les temps verbaux, les connecteurs, les adverbes et syntagmes adverbiaux de localisation temporelle.

Enfin, le résultat de ce travail est également la constitution d'un corpus nu, OWT-nu, constitué de 23 contes (environ 22 000 mots) et d'un corpus de référence annoté morphosyntaxiquement OWT-tag d'environ 7 500 mots qui seront mis à la disposition avec une licence Creative Commons d'autres chercheurs linguistes, littéraires, anthropologues… travaillant dans le domaine occitan.

Bibliographie

[ARM 06] ARMENTANO-OLLER, C., FORCADA M.-L., « Open-source machine translation between small languages: Catalan and Aranese Occitan », i*n Strategies for developing machine translation for minority languages (5th SALTMIL workshop on Minority Languages organized in conjunction with LREC)*, p. 51-54, 2006)

[BER 13] BERNHARD, D., LIGOZAT A.-L., « Es esch fàscht wie Ditsch, oder net? Étiquetage morpho-syntaxique de l'alsacien en passant par l'allemand », dans *Actes de TALARE 2013 : Traitement Automatique des Langues Régionales de France et d'Europe*, p. 209-220, 2013.

[BRA 16] BRAS M., VERGEZ-COURET, M., « BaTelÒc: a Text Base for the Occitan Language », In Vera Ferreira and Peter Bouda (eds.). *Language Documentation and Conservation in Europe*, pp. 133-149, Special Publication No. 9 of the Journal Language Documentation & Conservation, Honolulu: University of Hawaï Press, 2016.

[BRU 10] BRU J., « De l'oral à l'écrit : la rupture », *Port Acadie : revue interdisciplinaire en études acadiennes / Port Acadie: An Interdisciplinary Review in Acadian Studies*, numéro 16-17, p. 33-43, automne 2009, printemps 2010.

[CAR 05] Carruthers, J., *Oral Narration in Modern French, A Linguistic Analysis of Temporal Patterns*, Oxford: Legenda, 2005.

[CRY 01] Crystal, D., *Language and the Internet*, Cambridge, Cambridge University Press, 2001.

[FAN 08] FAN, R.-E., CHANG, K.-W., HSIEH, C.-J., WANG, X.-R. ET LIN, C.-J., « LIBLINEAR : A library for largelinear classification », in , Journal of Machine Learning Research, volume 9, p. 1871-1874, 2008.

[HAN 11] HANA, J., FELDMAN, A. AND AHARODNIK, K., « A low-budget tagger for Old Czech », in Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanitie (LaTeCH'11), p. 10-18, 2011.

[IDE 94] IDE, N., VÉRONIS, J., « MULTEXT (Multilingual Text Tools and Corpora) », in *Proceedings of the 14th International Conference on Computational Linguistics, COLING 94*, Kyoto, 1994.

[KOC 01] KOCH, P., OESTERREICHER, W. « Gesprochene Sprache und geschriebene Sprache/Langage parlé et langage écrit », in G. HOLTUS, M. METZELTIN, C. SCHMITT, (dir), *Lexicon der Romanistischen Linguistik*, volume I.2, pages 584–627, 2001.

[RAJ 97] RAJMAN, M., LECOMTE, J., PAROUBEK, P., « Format de description lexicale pour le français. Partie 2 : Description morpho-syntaxique », Rapp. Tech., EPFL & INaLF. GRACE GTR-3-2.1, 1997.

[REC 96] « 'ELM-FR: A typed French incarnation of the EAGLES-TS – Definition of Lexical Specification and Classification Guidelines », GSI-Erli, 1996.

[SAU 95] SAUZET, P., UBAUD, J., *Le verbe occitan. Lo vèrb occitan*. Aix-en-Provence:Édisud, 1995.

[SAU 06] SAUZET, P., *Conjugaison occitane. Savoir conjuguer en occitan (languedocien)*, IEO Edicions, 2006.

[SCH 13] SCHERRER Y., SAGOT B., « Lexicon induction and part-of-speech tagging of non-resourced and tools for closely related languages and language variants », dans *Actes de RANLP Workshop on Adaptation of language resources and tools for closely related languages and language variants*, p. 30-39, 2013.

[SIB 07] SIBILLE, J., « L'occitan, qu'es aquò ? », Langues et Cité : bulletin de l'observation des pratiques linguistiques, volume 10, numéro 2, 2007.

[TAC 13] TACKSTROM O., DAS D., PETROV S., MCDONALD R., NIVRE J., « Token and type constraints for cross-lingual part-of-speech tagging », dans *Actes de Transactions of the Association for Computational Linguistics*, p. 1-12, 2013.

[URI 13a] URIELI, A., VERGEZ-COURET, M. « Jochre, océrisation par apprentissage automatique : Etude comparée sur le yiddish et l'occitan », dans *Actes de la conférence TALN-RECITAL 2013 Volume 3 : Ateliers*, Les Sables d'Olonne: Université de Nantes, p. 221-234, 2013.

[URI 13b] URIELI A., *Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit*, Thèse de doctorat, Université de Toulouse 2 Le Mirail, 2013.

[VER 13] VERGEZ-COURET, M., « Tagging Occitan using French and Castillan Tree Tagger », in *Proceedings of "Less Resourced Languages, new technologies, new challenges and opportunities", 6th Language & Technology Conference*, Poznan, 2013.

[VER 14] VERGEZ-COURET, M., URIELI, A., « POS-tagging different varieties of Occitan with single-dialect ressources », in M. ZAMPIERI, L. TAN, N. LJUBEŠIĆ, J. TIEDEMANN (dir), *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, Dublin: Association for Computational Linguistics and Dublin City University, p. 21-29, 2014.

[VER 15a] VERGEZ-COURET, M., BERNHARD, D., URIELI, A., BRAS, M.; ERHART, P., HUCK D., « Numérisation et océrisation de textes pour les langues régionales : regards croisés sur l'occitan et l'alsacien », dans *10e colloque international ISKO France, Systèmes d'organisation des connaissances et humanités numériques*, Strasbourg, 2015.

[VER 15b] VERGEZ-COURET, M., URIELI, A., « Analyse morphosyntaxique de l'occitan languedocien : l'amitié entre un petit languedocien et un gros catalan, dans *Actes du Workshop Traitement Automatique des Langues Régionales de France et d'Europe 2*, Caen, 2015.

[VER 16] VERGEZ-COURET, M., « Description du lexique Loflòc », Rapport du projet ANR RESTAURE, 2016.

