Océrisation de textes pour les langues régionales. Regards croisés sur l'occitan et l'alsacien

Marianne Vergez-Couret, Delphine Bernhard, Assaf Urieli, Myriam Bras, Pascale Erhart, Dominique Huck

Océrisation de textes pour les langues régionales : regards croisés sur l'occitan et l'alsacien

VERGEZ-COURET Marianne
CLLE-ERSS, Université de Toulouse Jean Jaurès

BERNHARD Delphine
LiLPa, Université de Strasbourg

URIELI Assaf
CLLE-ERSS, Université de Toulouse Jean Jaurès
Joliciel Informatique, Foix

BRAS Myriam
CLLE-ERSS, Université de Toulouse Jean Jaurès

ERHART Pascale
LiLPa, Université de Strasbourg

HUCK Dominique
LiLPa, Université de Strasbourg

Résumé. Cet article détaille diverses expériences d'océrisation pour l'occitan et l'alsacien, en vue de constituer des corpus électroniques pour ces langues de France. Ces expériences visent à répondre à diverses questions : (i) quel outil d'océrisation, libre et gratuit, permet d'obtenir les meilleurs résultats ? (ii) est-il possible de réutiliser des modèles disponibles pour des langues proches déjà dotées de systèmes d'océrisation ? et (iii) est-il préférable d'utiliser des lexiques, éventuellement de petite taille, propres à la langue traitée ou des lexiques de plus grande taille pour des langues proches étymologiquement ? Nous comparons pour ce faire deux outils reposant sur des techniques d'apprentissage automatique supervisé, Jochre et Tesseract. Les résultats ne donnent pas d'avantage clair d'un outil sur l'autre, mais montrent que l'utilisation de modèles existants pour des langues proches est utile, dans la mesure où ces langues couvrent les caractères utilisés dans la langue cible. Par ailleurs, l'utilisation d'un lexique spécifique à la langue traitée, même de petite taille, est préférable à l'utilisation d'un lexique de grande taille pour une langue proche étymologiquement.

Mots-clés : océrisation, alsacien, occitan, lexiques, Jochre, Tesseract

1. Introduction

L'élaboration de nouveaux contenus numériques et le développement des technologies du langage sont aujourd'hui reconnus comme des moyens incontournables pour faire rayonner les langues et en particulier les langues régionales. Parmi les objectifs visés, il s'agit, pour les langues régionales, de dynamiser la recherche fondamentale, de favoriser le plurilinguisme et de développer des applications pour le public et les collectivités locales, avec la perspective globale de préservation et de valorisation du patrimoine culturel.

À l'heure actuelle, les langues régionales pâtissent du manque de ressources numériques. Dans le cadre du projet ANR RESTAURE, nous avons pour objectif de développer ressources et outils de base en traitement automatique des langues pour trois langues de France, dites "régionales" même si deux d'entre elles s'étendent au-delà du territoire français : l'alsacien, l'occitan et le picard. Nous souhaitons montrer comment les chercheurs travaillant sur des langues peu dotées très différentes peuvent collaborer, s'épauler pour les aspects techniques et développer une méthodologie commune. À notre sens, la toute première étape à viser est la constitution de corpus de langue écrite, de dictionnaires et de lexiques. Nous visons donc la constitution de corpus les plus représentatifs possible des usages de la langue, en rassemblant des œuvres écrites de différents genres (prose, théâtre, poésie, conte, presse...) et en accueillant la variation géolinguistique caractéristique des langues régionales (cf. section 2). Ces œuvres ne sont pas toujours disponibles au format numérique et la numérisation des fonds pour les langues régionales rencontre souvent une difficulté majeure qui est celle de l'absence de logiciel de reconnaissance optique de caractère (OCR) spécifiquement conçu pour la langue régionale en question.

Ainsi, il n'existe pas à notre connaissance d'outil d'océrisation pour l'alsacien, ce qui empêche la conversion en mode texte des ouvrages déjà numérisés. Les fonds occitans sont quant à eux en cours de numérisation par le CIRDOC (Centre Interrégional de Documentation Occitan) dans le cadre du projet Occitanica en partenariat avec la BNF (Bibliothèque Nationale de France). Les OCR qui ont été employés sont des OCR commerciaux (ABBYY FineReader et Adobe Acrobat Pro) qui ne

fournissent aucun détail technique et aucune évaluation sur le paramétrage pour l'occitan. Plusieurs études récentes (Holley, 2009 ; Boschetti *et al.*, 2009 ; Reynaert, 2008) ont d'ailleurs fait état d'une qualité d'océrisation moyenne à l'occasion de la numérisation de grands corpus bibliothécaires, même pour des œuvres en anglais. Malgré tout, l'océrisation est souvent considérée comme un problème résolu et peu ou plus évalué.

La première contribution de cet article est donc de comparer deux outils d'océrisation libres et gratuits, en utilisant une méthodologie commune pour l'alsacien et l'occitan. Nous avons choisi les outils Tesseract (Smith, 2007) et Jochre (Urieli & Vergez-Couret, 2013), qui utilisent tous deux des techniques d'apprentissage automatique supervisé. Ces techniques nécessitent la constitution de corpus annotés pour l'entraînement et l'évaluation, ainsi que des lexiques de la langue cible.

Dans le cas de Tesseract, il est possible de réutiliser des modèles entraînés pour des langues proches (français ainsi qu'allemand pour l'alsacien et catalan pour l'occitan) : notre deuxième contribution est donc de vérifier l'hypothèse selon laquelle des outils développés pour une langue proche peuvent être directement appliqués avec des taux de reconnaissance acceptables. Nous avons également testé, dans Jochre, l'utilisation de lexiques des langues étymologiquement proches (l'allemand pour l'alsacien et le catalan pour l'occitan).

Enfin, la dernière contribution est la constitution de modèles de reconnaissance optique de caractères pour l'alsacien et l'occitan, utilisables par les outils Tesseract et Jochre.

Nous allons dans un premier temps présenter les deux langues dont nous allons traiter : l'alsacien et l'occitan. Nous décrivons ensuite les outils d'océrisation utilisés, Jochre et Tesseract. La section suivante est consacrée à la présentation des ressouces utilisées (corpus et lexiques). Nous présentons dans la dernière section les résultats des évaluations de Tesseract et Jochre.

2. Présentation des deux langues régionales considérées : alsacien et occitan

Dans cet article, nous nous focalisons dans un premier temps sur l'alsacien et l'occitan, mais la méthodologie décrite sera amenée à être également appliquée au picard dans le cadre du projet ANR RESTAURE.

2.1. L'alsacien

Les dialectes parlés en Alsace, que l'on regroupe sous l'appellation « alsacien », appartiennent aux groupes alémanique et francique et se rapprochent de fait des dialectes parlés dans les régions limitrophes d'Allemagne et de Suisse. On retrouve des emprunts – essentiellement lexicaux – au français, mais la morphosyntaxe est très similaire à celle de l'allemand standard.

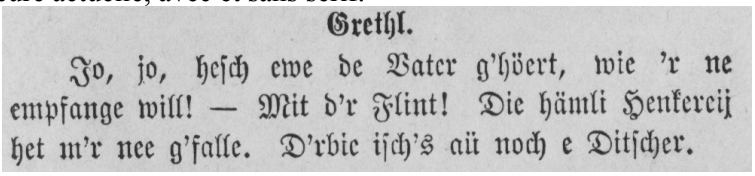
Les dialectes d'Alsace sont avant tout des langues parlées dans la vie quotidienne, et leur graphie n'est donc pas codifiée. L'OLCA (Office pour la Langue et la Culture de l'Alsace) précise d'ailleurs sur sa page Web de définition de la langue régionale : *«La langue normalisée, écrite et codifiée correspondante à nos dialectes est l'allemand standard.»*[1]

Il faut toutefois souligner que l'on trouve de nombreux exemples d'écrits en alsacien, répondant à des besoins à la fois culturels (pièces de théâtre, poésie, prose) et fonctionnels (lexiques et dictionnaires, méthodes d'apprentissage, sites web). Par ailleurs, certaines initiatives récentes se sont donné pour objectif de proposer une graphie propre à l'alsacien. On notera en particulier le système ORTHAL (Zeidler & Crévenat-Werner, 2008), qui se réfère à la norme graphique de l'allemand standard tout en permettant la transcription des spécificités des dialectes alsaciens.

Si les écrits existent en alsacien, leur numérisation est compliquée non seulement par la variété des formes de scripturalisation employées, mais aussi par l'utilisation majoritaire de l'écriture « Fraktur », généralement appelée écriture gothique, jusque dans l'immédiat après Première Guerre mondiale (voir Figure 1 ci-dessous). Cela a pour conséquence que la

[1] http://www.olcalsace.org/fr/observer-et-veiller/definition-de-la-langue-regionale. Auteurs : Adrien Finck, Frédéric Hartweg, Raymond Matzen, Marthe Philip.

reconnaissance optique de caractères pour les textes imprimés en alsacien nécessite deux modèles différents : un modèle « Fraktur » pour les textes les plus anciens, et un modèle pour les polices de caractères plus utilisées à l'heure actuelle, avec et sans sérif.


Grethl.

Jo, jo, heſch ewe de Vater g'hört, wie 'r ne empfange will! — Mit d'r Flint! Die hämli Henkereij het m'r nee g'falle. D'rbie iſch's aü noch e Ditſcher.

Figure 1 : exemple de caractères « Fraktur » en alsacien. Extrait de *D'r Herr Maire*, *Lustspiel in dreij Akt*, de Gustave STOSKOPF, Strassburg 1898, Schlesier & Schweikhardt.

2.2. L'occitan

L'occitan est une langue romane, parlée dans le sud de la France, le Val d'Aran en Espagne et dans 12 vallées alpines d'Italie. Cette langue ne dispose pas d'une variété standard. Six variantes sont généralement distinguées : l'auvergnat, le gascon, le languedocien, le limousin, le provençal et le vivaro-alpin (Bec, 1995), chaque variante connaissant une variation interne. En ce qui concerne les graphies de l'occitan, la langue est écrite avec un alphabet latin depuis le moyen-âge. Mais, depuis la production des troubadours, plusieurs graphies ont été adoptées. Aujourd'hui, on distingue une convention graphique dite classique, proche de celle des troubadours, une graphie dite mistralienne, issue de la renaissance littéraire impulsée par le Félibrige au XIX^{e} siècle, ainsi que plusieurs autres conventions graphiques plus ou moins individuelles (Sibille, 2007).

La numérisation de textes occitans a été entreprise par le CIRDOC et par l'association CIEL d'OC dans le cadre de la constitution de bibliothèques virtuelles, rendant l'existence d'un OCR spécifique et performant nécessaire. Par ailleurs, la constitution de la base textuelle BaTelÒc (Bras et Thomas, 2011), réunissant aujourd'hui un corpus d'environ trois millions de mots issus d'œuvres contemporaines existant au format numérique, va devoir intégrer des œuvres plus anciennes qu'il faudra numériser et océriser de façon efficace. La matière écrite occitane à rassembler étant estimée à plusieurs centaines de millions de mots, la

création d'un OCR performant est donc un enjeu majeur pour la création de ressources numériques pour cette langue.

3. Outils d'océrisation

3.1. Jochre

Le logiciel Jochre (Java Optical CHaracter REcogition) est un logiciel OCR libre développé par Assaf Urieli, reposant sur des techniques d'apprentissage automatique supervisé. L'analyse de Jochre s'effectue en trois étapes :

Segmentation des images en paragraphes, lignes, mots et « formes ».

La segmentation utilise des techniques statistiques *ad hoc* adaptées à chaque tâche (détection de l'orientation, suppression des petites taches, … voir Figure 2).

Figure 2 : exemple de segmentation d'un texte en yiddish.

Reconnaissance des lettres

La reconnaissance des lettres s'applique à chaque forme retrouvée lors de l'étape précédente. Elle utilise des techniques d'apprentissage automatique supervisé, et se divise donc en deux étapes distinctes : l'entraînement et l'analyse. Lors de l'entraînement, l'utilisateur charge des images scannées vers un logiciel web, et attribue la bonne lettre à chaque forme. Ces images annotées en lettres constituent le corpus d'apprentissage. Un modèle statistique est ensuite entraîné à partir de ce corpus, et va servir à la reconnaissance automatique des lettres dans une nouvelle image. Pour pouvoir entraîner ce modèle, chaque forme annotée avec une lettre dans les images scannées est décrite automatiquement à l'aide d'une liste de descripteurs (*feature* en anglais). Ces descripteurs

peuvent concerner des informations graphiques (*ex.* la hauteur ou largeur de la forme) ainsi que des informations tirées du contexte (*ex.* les n-grammes de lettres précédentes). Ensuite, un classifieur robuste analyse toutes les formes annotées pour savoir quelles lettres sont favorisées par chaque résultat de descripteur (*ex.* si la forme est étroite, la lettre correspondante sera plus probablement un "i" qu'un "m"). Suivant un algorithme itératif complexe, le classifieur attribue un poids relatif à chaque résultat de descripteur pour chaque lettre, et stocke ces poids dans le *modèle statistique*. Quand le classifieur doit analyser une forme à annoter automatiquement dans une nouvelle image, il décrit la forme avec les mêmes descripteurs, et utilise le modèle statistique pour générer une distribution de probabilités pour chaque lettre de l'alphabet. C'est ainsi qu'il peut proposer les *n* analyses les plus probables pour chaque mot. Jochre propose deux classifieurs robustes pour la construction des modèles : l'entropie maximale ou MaxEnt (Rathnaparkhi, 1998) et le SVM linéaire (Fan *et al.*, 2008).

Correction des mots à l'aide du lexique

Pendant la correction, Jochre utilise un lexique pour reclasser les analyses, donnant un poids plus important aux mots trouvés dans le lexique. Pour ce faire, il diminue le score initial d'un mot inconnu en le multipliant par un coefficient de réduction. Dans la Figure 3 ci-dessous, la mauvaise analyse "acordot" avait le score initial le plus élevé, mais suite à l'application de la coefficient de réduction de 0,5 aux mots inconnus dans le lexique, l'analyse correcte "acordat" est montée en tête.

	score initial	connu ?	score ajusté
acordot	**72,0 %**	non (x 0,5)	36,0 %
acordat	70,1 %	**oui** (x 1,0)	**70,1 %**
acordet	64,3 %	non (x 0,5)	32,2 %

Figure 3 : exemple de reclassement des analyses en occitan.

3.2. Tesseract

Tesseract est un outil de reconnaissance optique de caractères développé à l'origine à HP Labs entre 1984 et 1994 (Smith, 2007) et diffusé sous forme de projet open source en 2005 (disponible à https://github.com/tesseract-ocr). À l'heure actuelle, plus de 60 langues

sont disponibles. Les résultats obtenus par Tesseract sont très bons pour les scripts latins : pour l'anglais, le taux d'erreur par caractère est de 0.5 % et le taux d'erreur par mot de 3.72 % (Smith et al., 2009). Tesseract peut être appliqué à des textes multilingues (plusieurs langues dans la même image) en combinant plusieurs modèles de reconnaissance. Tout comme Jochre, Tesseract utilise des techniques d'apprentissage automatique supervisé pour lesquelles les ressources présentées dans la section suivante ont été créées.

4. Ressources pour l'alsacien et l'occitan et entraînement des outils

Pour chaque nouvelle langue à intégrer dans Jochre ou Tesseract, il faut créer des ressources : corpus annotés (exemples des sorties souhaitées) et des lexiques, c'est-à-dire des listes de formes fléchies (représentant autant que possible la variété graphique et dialectale).

4.1. Préparation des corpus et lexiques

Les corpus de pages scannées et annotées, cf. Figure 4 ci-dessous, sont divisés en deux. Une partie est exploitée pour entraîner Jochre et Tesseract et l'autre partie est conservée pour l'évaluer.

Figure 4. Interface d'annotation.

Les corpus sont constitués de plusieurs pages (> 60) de livres variés de sorte à diversifier les types, les tailles et les styles (gras, italique… de

police (environ 6 pages par livre). Les contenus des corpus sont présentés dans la Table 1.

	Corpus alsacien	Corpus occitan
Nombre d'ouvrages	11 entraînement : 7 test : 4	11 entraînement : 7 test : 4
Genres	prose, pièce de théâtre, poésie	roman, préface
Variétés représentées	Diverses	Deux dialectes : languedocien et gascon
Polices	sérif, sans sérif, italique	sérif, italique
Années d'édition	1890-2000	1960-2000
Nombre de pages	68 entraînement : 43 test : 25	86 entraînement : 53 test : 33
Nombre de mots	13 900 entraînement : 9 000 test : 4 900	20 400 entraînement : 13 900 test : 6 500
Nombre de caractères	57 300 entraînement : 36 150 test : 21 150	84 000 entraînement : 54 700 test : 29 300

TABLE 1 – Caractéristiques des corpus utilisés pour l'entraînement et l'évaluation.

Les corpus ont été annotés manuellement à l'aide de l'outil JochreWeb puis corrigés manuellement après une première analyse automatique du corpus avec Jochre en les confrontant là où l'annotation manuelle et l'annotation automatique différaient.

Les lexiques sont des listes de mots constituées à partir des entrées de plusieurs dictionnaires et d'œuvres déjà disponibles au format numérique. Les logiciels d'OCR peuvent utiliser ces listes pour donner un poids plus important ou privilégier les mots présents dans ces lexiques.

Lexiques pour l'alsacien

Pour l'alsacien, nous avons utilisé divers lexiques disponibles sur le Web et utilisés précédemment pour d'autres travaux (Bernhard et Steiblé, 2015) :

- les lexiques produits par l'OLCA[2] (Office pour la Langue et la Culture d'Alsace). 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 généralement des variantes pour les départements alsaciens du Bas-Rhin et du Haut-Rhin ;
- un lexique extrait d'une page utilisateur du Wiktionnaire[3];
- un lexique bilingue disponible sur la page Web d'une association locale, l'ACPA[4].

Ces lexiques contiennent pour l'essentiel des formes lemmatisées. Nous les avons donc complétés avec les formes extraites de divers corpus, afin d'enrichir le lexique final avec des formes fléchies. Les corpus utilisés sont les suivants :

- Un ensemble de 65 chroniques publiées par Raymond Matzen dans le quotidien local "Les Dernières Nouvelles d'Alsace" ;
- Un ensemble de transcriptions d'émissions télévisées en langue régionale réalisées dans le cadre de la thèse de Pascale Erhart (Erhart, 2012) ;
- Un corpus parallèle français – alsacien comportant des traductions du français vers l'alsacien réalisées par l'OLCA ;
- Un extrait du « Dictionnaire comparatif multilingue » de Paul Adolf (Adolf, 2006), qui comprend notamment des phrases en alsacien qui illustrent les différentes entrées.

Le lexique final extrait à partir des lexiques et des corpus comprend 56 795 formes.

---

[2] http://www.olcalsace.org/

[3] http://fr.wiktionary.org/wiki/Utilisateur:Laurent_Bouvier/alsacien-fran\%C3\%A7ais

[4] Compilé par André Nisslé, http://culture.alsace.pagesperso-orange.fr/dictionnaire_alsacien.htm

En plus de ce lexique de formes alsaciennes, nous avons utilisé un lexique de mots allemands téléchargé sur le site du projet *Wortschatz* de l'université de Leipzig[5] et qui comprend plus d'1 million de formes issues de la Wikipédia en allemand. L'hypothèse sous-jacente est qu'un certain nombre de mots sont écrits avec leur graphie allemande dans les documents alsaciens.

Pour vérifier l'hypothèse de la proximité entre alsacien et allemand, nous avons effectué une comparaison des caractères que l'on trouve dans le corpus d'entraînement et de test, par rapport aux caractères français et allemands. Les résultats figurent dans la Table 2. On constate que les caractères allemands couvrent plus de 98% du corpus. Cela étant, la couverture est encore meilleure pour les caractères français, car elle est de plus de 99%.

Nombre de caractères alphabétiques	57 332
Nombre de caractères allemands	56 372 (98,33%)
Nombre de caractères français	56 769 (99,02%)
Nombre de caractères français non allemands	960 (1,67%)
Nombre de caractères allemands non français	563 (0,98%)

TABLE 2 – Caractères alphabétiques dans le corpus alsacien.

Nous avons également fait un décompte du nombre d'occurrences des caractères spécifiques à chaque alphabet (voir Table 3).

Caractères spécifiquement allemands	
ä	419 (0,73%)
Ä	5 (0,01%)
ö	83 (0,14%)
ß	56 (0,10%)

[5] http://corpora2.informatik.uni-leipzig.de/

Caractères spécifiquement français	
à	651 (1,14%)
À	22 (0,04%)
â	75 (0,13%)
ç	4 (0,01%)
Ç	1 (0,002%)
é	154 (0,27%)
É	1 (0,002%)
è	21 (0,04%)
ê	1 (0,002%)
î	23 (0,04%)
ï	4 (0,01%)
ô	2 (0,003%)
ù	1 (0,002%)

TABLE 3 – Nombre d'occurrences et fréquence des caractères spécifiques à chaque alphabet.

Le caractère spécifiquement allemand le plus représenté est "ä", qui apparaît d'ailleurs dans tous les documents du corpus sans exception. Pour le français, c'est "à"[6], qui même s'il a un grand nombre d'occurrences ne se trouve pas dans tous les documents. Le caractère allemand "ß" n'apparaît que dans deux documents. On constate également qu'il y a plus de caractères français différents qui apparaissent dans les documents alsaciens que de caractères allemands (13 contre 4).

Lexiques pour l'occitan

[6] Ce caractère est utilisé dans la majorité des cas pour rendre compte d'un son typiquement alsacien (a suédois, c'est-à-dire [ɔ], a vélaire).

Pour l'occitan, nous avons tout d'abord extrait le lexique des 60 œuvres issues (de 29 auteurs différents) du corpus de textes rassemblés dans le cadre du projet BaTelÒc pour les dialectes languedociens et gascons (correspondant aux deux dialectes présents dans les corpus d'entraînement et d'évaluation des OCR). Ce lexique comprend 432 534 formes.

Nous avons également utilisé des dictionnaires et lexiques disponibles au format numérique :

- Dictionnaire Français/Occitan Gascon Toulousain de Nicolau Rei Bèthvéder, 2004, IEO Edicions
- Dictionnaire Français/Occitan de Cristian Laus, 2004, IEO/IDECO
- Dictionnaire Français/Occitan (Gascon) de Miquèu Grosclaude, Gilabèrt Nariòo e Patric Guilhemjoan, 2007, Per Noste Edicions
- Une partie du Verb'Òc (formes conjuguées des verbes en languedocien) mis en ligne par le *Congrès permanent de la lenga occitana*[7]
- Une liste de noms propres extraite du lexique occitan d'Apertium, outil de traduction automatique de langues étymologiquement proches[8]

Nous disposons ainsi de 71 589 entrées pour le lexique gascon et de 255 399 entrées pour le lexique languedocien. Le nombre d'entrées cumulées pour les dictionnaires et lexiques s'élève à 300 563. En tout, nous disposons pour l'occitan d'un lexique de 432 537 formes.

Afin de tester l'hypothèse qu'il est possible de tirer partie des ressources d'une langue étymologiquement proche, nous allons utilisé le lexique catalan d'Apertium comportant 715 166 formes fléchies.

L'examen de la fréquence des caractères français et catalan dans notre corpus occitan permet de vérifier l'hypothèse de la proximité de l'occitan avec les deux langues. On constate, dans la Table 4 que les caractères français couvrent plus de 98% du corpus tandis que les caractères catalans avoisinent les 100%.

[7] http://www.locongres.org
[8] http://sourceforge.net/projects/apertium/files/?source=navbar

Nombre de caractères alphabétiques	83 997
Nombre de caractères français	82 759 (98,53%)
Nombre de caractères catalans	83 989 (99,99%)
Nombre de caractères français non catalans	8 (0,009%)
Nombre de caractères catalans non français	1 238 (1,47%)

TABLE 4 – Caractères alphabétiques dans le corpus occitan.

Nous avons également fait un décompte du nombre d'occurrences des caractères spécifiques à chaque alphabet (voir Table 5) qui confirme la très grande proximité de l'alphabet occitan avec celui du catalan.

Caractères spécifiquement français	Nombre d'occurrences / pourcentage
Ç	7 (0,008%)
œ	1 (0,0012%)
Caractères spécifiquement catalans	
ò	933 (1,11%)
í	118 (0,14%)
ó	100 (0,12%)
á	68 (0,08%)
ú	14 (0,017%)
Ò	5 (0,06%)

TABLE 5 – Fréquence des caractères spécifiques à chaque alphabet.

4.2. Création de modèles pour l'alsacien et l'occitan

Nous avons utilisé et entraîné les deux outils de reconnaissance de caractères présentés plus haut, Tesseract et Jochre, afin de comparer leurs performances.

Nous avons utilisé Tesseract de deux manières différentes : (i) application de modèles existants pour des langues proches : français, ainsi que catalan pour l'occitan et allemand pour l'alsacien et (ii) entraînement de nouveaux modèles pour l'occitan et l'alsacien à partir de nos corpus.

Pour appliquer les modèles fournis par Tesseract, nous avons utilisé le logiciel *gImageReader*[9], développé par Sandro Mani. Le logiciel constitue une interface graphique pour Tesseract : il permet de charger et de visualiser des images, puis de reconnaître le texte à l'aide de Tesseract. *gImageReader* permet ensuite de reconstituer les paragraphes en supprimant les sauts de ligne superflus.

L'entraînement de Tesseract se fait à l'aide du logiciel *jTessBoxEditor*[10]. Afin de faciliter le processus, l'apprentissage ne se fait pas à partir d'images scannées "réelles" mais à partir d'images générées automatiquement à partir des textes d'entraînement corrigés au format texte brut. Ceci permet de faire automatiquement la correspondance entre les caractères dans l'image et le résultat souhaité pour l'OCR. Le texte d'entraînement doit contenir plusieurs exemplaires de chaque caractère : 5 au moins pour les caractères rares et au moins 20 pour les caractères les plus fréquents. Il est possible d'utiliser un texte "artificiel" (suite de caractères qui n'a pas nécessairement de sens), mais nous avons fait le choix d'utiliser notre corpus d'entraînement, ce qui nous permet de comparer Jochre et Tesseract avec des données d'entraînement similaires. La génération des images à partir de textes bruts se fait en fonction de différents paramètres : taille de police (nous avons utilisé 36pt), type de police (nous avons utilisé Arial et Times New Roman, normal et italique). Il est également possible d'ajouter artificiellement du bruit à l'image, ainsi que d'augmenter l'espacement entre les caractères mais nous n'avons pas utilisé ces possibilités. jTessBoxEditor permet également de lancer l'entraînement de Tesseract. Tesseract utilise deux listes de mots pour l'entraînement. La première contient les mots fréquents : nous avons utilisé les 1 000 mots les plus fréquents de notre lexique, en prenant en

[9] https://github.com/manisandro/gImageReader
[10] http://vietocr.sourceforge.net/training.html

compte les occurrences dans les corpus et les lexiques utilisés. La seconde contient l'ensemble des mots du lexique.

5. Évaluation

5.1. Mesures et outils d'évaluation

Nous avons appliqué différentes mesures et outils d'évaluation, afin d'estimer l'exactitude et le taux d'erreurs pour la reconnaissance des caractères et des mots :

- *ocropus-econf* : script d'évaluation fourni par le projet ocropy[11]. Le script donne le pourcentage d'erreurs pour les caractères.
- *wdiff* : programme basé sur diff qui permet de comparer les mots apparaissant dans deux fichiers et ainsi de compter le nombre de mots qui se trouvent à la fois dans le résultat de l'OCR et le fichier de référence corrigé manuellement (mots communs). La mesure donne le pourcentage de mots communs par rapport au nombre total de mots dans le fichier de référence.

Ces deux outils ont notamment été utilisés par Garcia-Fernandez et al. (2014) pour l'évaluation de l'OCR. Comme (Garcia-Fernandez et al., 2014), nous avons également normalisé certains caractères, comme les différents types d'apostrophes, afin d'éviter un impact négatif trop important sur la mesure de performance.

5.2. Résultats de l'évaluation pour l'alsacien

Nous présentons tout d'abord les résultats obtenus par Tesseract, dans différentes configurations :

1. TESS-DEU : application du modèle allemand fourni sur le site de Tesseract[12]
2. TESS-FRA : application du modèle français fourni sur le site de Tesseract

[11] https://github.com/tmbdev/ocropy

[12] https://code.google.com/p/tesseract-ocr/downloads/list

3. TESS-GSW : entraînement de Tesseract à partir du corpus d'entraînement, selon la procédure décrite précédemment

Les autres configurations correspondent à l'application conjointe d'au moins deux modèles (allemand, alsacien ou français) en utilisant la fonctionnalité "multilingue" de Tesseract. La Table 6 donne les résultats obtenus pour les différentes configurations :

	Pourcentage d'erreurs (caractères)	**Pourcentage de mots communs (après normalisation)**
TESS-DEU	2,33%	91,25%
TESS-FRA	2,42%	91,03%
TESS-GSW	3,20%	87,39%
TESS-DEU+GSW	2,03%	92,54%
TESS-FR+GSW	2,19%	91,27%
TESS-FR+DEU	1,68%	94,27%
TESS-DEU+FR+GSW	**1,63%**	**94,46%**

TABLE 6 – Évaluation de Tesseract pour l'alsacien.

Les résultats montrent que les modèles fournis par le projet Tesseract pour l'allemand et le français obtiennent de meilleurs résultats que le modèle alsacien. L'entraînement a vraisemblablement été réalisé pour une plus grande variété de polices, ce qui pourrait expliquer les meilleurs résultats. Cela étant, les résultats obtenus pour l'entraînement sur le corpus alsacien restent acceptables, compte tenu de la petite taille du corpus d'entraînement. Les meilleurs résultats sont obtenus pour la combinaison de tous les modèles. Il faut souligner l'apport du modèle français, qui permet de passer de 91,25% de mots communs à 94,27% par rapport à l'utilisation du modèle allemand seul. Comme nous l'avons montré précédemment, les caractères spécifiquement français, que l'on ne retrouve pas en allemand, sont fréquemment utilisés en alsacien. D'une manière générale, les problèmes de reconnaissance constatés concernent des cas classiques de confusion entre caractères similaires graphiquement

(*c/e*, *i/l*, *q/g*, *W/w*, *0/o*, etc.), ainsi que des confusions entre caractères accentués, en particulier des caractères spécifiquement français (*ê/é*, *ä/à*, *à/a*, *i/î*, *a/ä*, etc.).

Nous présentons maintenant les résultats obtenus par Jochre.

Nous avons testé deux modèles d'apprentissage différents : MaxEnt et SVM. Suivant Urieli et Vergez-Couret (2013), nous avons fait varier le coefficient de réduction de score pour les mots inconnus, en le faisant varier de 0,1 à 1,0. Les meilleurs résultats sont obtenus pour le modèle SVM avec un coefficient de réduction de 0,6 et pour le modèle MaxEnt avec un coefficient de 0,8.

La Table 8 présente les résultats d'évaluation en utilisant le même corpus de test et les mêmes mesures que pour l'évaluation de Tesseract. Nous faisons également varier les ressources pour l'analyse : nolex (sans lexique), gsw (lexique alsacien), deu (lexique allemand), deu+gsw (combinaison des lexiques alsacien et allemand). Globalement l'utilisation de lexiques a un impact positif sur les performances, en particulier l'utilisation du lexique alsacien.

	Pourcentage d'erreurs (caractères)	**Pourcentage de mots communs (après normalisation)**
MaxEnt-nolex	8,49%	75,62%
MaxEnt-gsw	7,56%	79,20%
MaxEnt-deu	7,69%	78,85%
MaxEnt-deu+gsw	7,44%	**79,65%**
SVM-nolex	8,39%	73,72%
SVM-gsw	7,06%	79,59%
SVM-deu	7,25%	78,68%
SVM-deu+gsw	**7,01%**	79,63%

TABLE 8 – Évaluation de Jochre pour l'alsacien.

L'analyse des résultats montre que Jochre confond plus fréquemment des caractères fréquents et non accentués que Tesseract : *e/c*, *a/u*, *n/u*, *H/B*, *t/l*, *l/I*, *f/l*, etc., d'où les performances plus faibles.

5.3. Résultats de l'évaluation pour l'occitan

Nous présentons dans la table 9 les résultats obtenus par Tesseract pour des configurations similaires à celles précédemment présentées pour l'alsacien :

1. TESS-CAT : application du modèle catalan fourni sur le site de Tesseract
2. TESS-FR : application du modèle français fourni sur le site de Tesseract
3. TESS-FR+CAT: application des modèles français et catalan en utilisant la fonctionnalité "multilingue" de Tesseract (afin de couvrir tous les caractères présents en occitan)
4. TESS-OCC : entraînement de Tesseract à partir de la moitié du corpus d'entraînement (il a été nécessaire de couper en deux le corpus en raison de difficultés techniques), selon la procédure décrite précédemment
5. TESS-OCC+CAT: application des modèles occitan et catalan en utilisant la fonctionnalité multilingue de Tesseract
6. TESS-OCC+CAT+FR : application des modèles occitan, catalan et français en utilisant la fonctionnalité multilingue de Tesseract

Tous les résultats obtenus se situent sur une échelle variant de 72,18% (TESS-FRA) à 79,50% (TESS-CAT+FRA+OCC). Pour les modèles unilingues (CAT, FRA, OCC), les meilleurs résultats sont obtenus avec le modèle catalan (75,04%) et non pas avec le modèle occitan. Comme vu précédemment, l'entraînement a été réalisé avec un petit corpus et pour un nombre restreint de polices. Avec le modèle français, les résultats sont moins bons mais la combinaison des modèles FRA+CAT (qui permet de couvrir tous les caractères employés en occitan) permet une nette amélioration (78,22%). Ces résultats renforcent l'hypothèse selon laquelle il est possible d'obtenir des résultats sans avoir de modèles spécifiques à la langue peu dotée. Le meilleur résultat (79,50%) est

obtenu avec le modèle multilingue (CAT+FRA+OCC), montrant tout de même l'apport du modèle occitan.

	Pourcentage d'erreurs (caractères)	**Pourcentage de mots communs (après normalisation)**
TESS-CAT	7,81%	75,04%
TESS-FRA	8.46%	72,18%
TESS-FRA+CAT	6,64%	78,22%
TESS-OCC	7,29%	73,33%
TESS-OCC+CAT	6,51%	74,95%
TESS-CAT+FRA+OCC	**5,76%**	**79,50%**

TABLE 9 – Évaluation de Tesseract pour l'occitan.

Notons que les moyennes proposées dans cette table cachent une grande disparité dans les résultats obtenus pour les 4 documents du corpus de test, les résultats variant d'une vingtaine de points pour toutes les configurations. Ici, la très moyenne qualité de la numérisation pour deux documents a un impact négatif sur les résultats. Il sera intéressant de comparer pour ces documents les résultats de Jochre afin de voir si les mêmes difficultés sont rencontrées.

Concernant les caractères fréquemment confondus, nous relevons des cas classiques de confusion entre caractères similaires graphiquement (*c/e*, *I/l*, etc.) et les variantes majuscules/minuscules d'un même caractère (C/c, V/v, etc.), ainsi que des confusions entre caractères accentués (*í/i*, *e/è*, *é/è*, *ò/è*, etc.).

Nous présentons maintenant les résultats obtenus par Jochre. Dans (Urieli et Vergez-Couret, 2013), nous avions fait varier le coefficient de réduction de score pour les mots inconnus de 0,1 à 1,0. Pour le modèle MaxEnt, les meilleurs résultats sont obtenus avec un coefficient de réduction de 0,75 (exactitude supérieure à 80%). Dans cet article, nous avions également mis en avant l'amélioration apportée par l'utilisation

des lexiques. Les résultats montrent l'importance de disposer de lexiques de taille importante. Néanmoins, certains des résultats indiquaient également l'importance de la précision des lexiques, au niveau du dialecte.

Pour le présent article, nous avons testé deux modèles d'apprentissage différents : MaxEnt et SVM. Les résultats sont présentés dans la Table 9 avec les mêmes corpus que pour l'évaluation de Tesseract, en utilisant le coefficient de réduction : 0,75 pour MaxEnt et 0,75 pour SVM.

Nous faisons également varier les ressources pour l'analyse : nolex (sans lexique), occ (lexique occitan), cat (lexique catalan), cat+occ (combinaison des lexiques catalans et occitans).

	Pourcentage d'erreurs (caractères)	**Pourcentage de mots communs (après normalisation)**
MaxEnt-nolex	4,50%	81,90%
MaxEnt-occ	3,29%	88,43%
MaxEnt-cat	3,53%	87,34%
MaxEnt-cat+occ	3,24%	88,42%
SVM-nolex	3,39%	88,29%
SVM-occ	**3,07%**	**89,36%**
SVM-cat	3,39%	88,27%
SVM-cat+occ	3,11%	89,23%

TABLE 10 – Évaluation de Jochre pour l'occitan.

Comme vu précédemment avec l'évaluation de Tesseract, ces résultats cachent une grande variation d'un document à l'autre. Le pourcentage d'erreur au niveau du caractère pour SVM-occ varie de 1,6% à 4% et le pourcentage de mots communs de 86,6% à 92,8%. Les variations d'un document à l'autre sont moins importantes avec Jochre. Globalement, les résultats sont meilleurs que ceux obtenus avec Tesseract. Mais surtout,

ces résultats confirment l'importance de posséder un lexique de grande taille pour la langue analysée. Dans ce cas, il est même inutile d'utiliser des ressources de langues proches.

6. Conclusion et perspectives

Nous pouvons tirer plusieurs conclusions méthodologiques de ces expériences. Premièrement, il est possible d'obtenir des sorties d'OCR de qualité satisfaisante en utilisant des modèles développés pour des langues proches. Cela étant, la proximité linguistique qui semble la plus intéressante ici n'est pas seulement celle de famille linguistique, mais également celle de la proximité graphique : il est donc préférable de choisir plusieurs modèles couvrant l'ensemble des caractères utilisés dans la langue à analyser. Ainsi, l'alsacien, qui est graphié avec un grand nombre de caractères français, bénéficie de l'utilisation d'un modèle de reconnaissance de caractères développé pour le français, même si le français et les dialectes alsaciens appartiennent à des familles linguistiques différentes. Il semble donc utile, pour guider le choix des modèles à utiliser, de tout d'abord faire une étude des caractères utilisés dans la langue à traiter afin de déterminer les langues proches dans lesquelles on retrouve ces caractères. La deuxième conclusion est que l'utilisation de lexiques spécifiques améliore les résultats. Toutefois, cela n'est pas toujours simple à mettre en œuvre de manière satisfaisante pour des langues dont la graphie n'est pas normée et pour lesquelles il n'est donc pas possible d'obtenir un lexique exhaustif. À cela s'ajoute le manque de ressources disponibles.

La comparaison de Jochre et Tesseract ne donne pas d'avantage clair d'un outil sur l'autre : si Tesseract obtient de meilleurs résultats que Jochre pour l'alsacien, on constate la tendance inverse pour l'occitan. Diverses raisons peuvent être avancées. Tout d'abord, le corpus d'entraînement utilisé pour l'alsacien pour Jochre est plus petit que celui utilisé pour l'occitan. Par ailleurs, il mélange des textes en police avec et sans empattement, tandis que le corpus d'entraînement utilisé pour l'occitan est plus stable (uniquement des caractères sérif, avec empattement). Il faudrait donc augmenter la taille du corpus d'entraînement pour Jochre en alsacien pour obtenir des conditions

d'entraînement similaires à celles observées pour l'occitan. Pour ce qui est de Tesseract, les mêmes conclusions s'imposent : il faudrait améliorer le corpus d'entraînement, soit en augmentant sa taille, soit en le complétant avec des fichiers "réels" (c'est-à-dire de vrais documents scannés et non des images générées à partir de textes).

Au-delà du lexique externe, il reste d'autres pistes d'amélioration possibles. Une première possibilité serait d'ajouter une étape de post-traitement, permettant de corriger des erreurs fréquemment observées, ou encore des erreurs spécifiques et récurrentes dans certaines œuvres (par exemple, les noms de personnages qui se répètent dans une pièce de théâtre).

À l'issue de ces expériences, nous avons dans un premier temps doté l'alsacien et l'occitan de ressources, corpus annotés et lexiques, pour entraîner et évaluer deux logiciels OCR, ce qui a abouti à la création de modèles de reconnaissance optique de caractères pour ces deux langues. Dans un second temps, la méthodologie mise au jour sera appliquée à la langue picarde dans le cadre du projet ANR RESTAURE et pourra être également appliquée à toutes langues qui souhaitent disposer de ces ressources.

7. Remerciements

Nous remercions l'OLCA, André Nisslé et Paul Adolf ainsi que Lo Congrès Permanent de la Lenga Occitana et l'IEO du Tarn pour nous avoir donné accès à leurs ressources. Nous remercions également Laura Katz pour son aide lors de la correction du corpus d'apprentissage de l'alsacien, Anne-Laure Ligozat pour nous avoir fourni les scripts d'évaluation et suggéré l'utilisation de Tesseract et Basilio Calderone pour son soutien.

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

Références

Adolf, P. (2006). Dictionnaire comparatif multilingue: français-allemand-alsacien-anglais., Strasbourg, France, Midgard, 2006, 373 p.

Bernhard, D. et Steiblé, L. (2015). Quand l'oral se fait entendre à l'écrit : alignement de lexiques en l'absence de normalisation graphique. In Actes de l'atelier sur le Traitement Automatique des Langues Régionales de France et d'Europe, Caen, France.

Boschetti, F., Romanello, M., Babeu, A., Bamman, D., et Crane, G. (2009). Improving OCR accuracy for classical critical editions. In Research and Avdanced Technology for Digital Libraries, pp. 156-167. Springer Berlin Heidelberg.

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'Association Internationale d'Etudes Occitanes, Aix-la-Chapelle, 24-31 août 2008, Aachen, Shaker.

Breuel, T. M. (2008) The OCRopus open source OCR system. Proceedings IS&T/SPIE 20th Annual Symposium.

Erhart, P. (2012). Les dialectes dans les médias : quelle image de l'Alsace véhiculent-ils dans les émissions de la télévision régionale ?, Université de Strasbourg Consultable à http://www.theses.fr/167563386 [Accédé le 7 mai 2013].

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R. et Lin, C.-J. (2008). LIBLINEAR : A library for largelinear classification, Journal of Machine Learning Research, vol. 9, p. 1871-1874, 2008.

Garcia-Fernandez, A., Ligozat, A.-L., et Vilnat, A. (2014). Construction and annotation of a french folkstale corpus. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland.

Holley, R. (2009). How good can it get ? Analysing and improving OCR accuracy in large scale historic newspaper digitisation programs. D-Lib Magazine 15, no. ¾.

Ratnaparkhi, A. (1998). Maximum entropy models for natural language ambiguity resolution. Thèse, University of Pennsylvania, 1998.

Reynart, M. (2008). Non-interactive OCR post-correction for giga-scale digitization projets. Computational Linguistics and Intelligent text Processing. Springer Berlin. 617-630.

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

Smith, R. (2007). An overview of the Tesseract OCR engine. Ninth International Conference on Document Analysis and Recognition. ICDAR 2007. Vol. 2. IEEE.

Smith, R., Antonova D., et Lee D.-S. (2009). Adapting the Tesseract open source OCR engine for multilingual OCR. In Proceedings of the International Workshop on Multilingual OCR, ACM.

Urieli, A. et Vergez-Couret, M. (2013). Jochre, océrisation par apprentissage automatique : Etude comparée sur le yiddish et l'occitan, Actes de la conférence TALN-RECITAL 2013 Volume 3 : Ateliers (Les Sables d'Olonne: Université de Nantes), 221-234.

Zeidler, E. et Crévenat-Werner D. (2008). Orthographe alsacienne: bien écrire l'alsacien de Wissembourg à Ferrette., Colmar, France, J. Do Bentzinger, 2008, 143 p.