Jochre, océrisation par apprentissage automatique : étude comparée sur le yiddish et l'occitan

Assaf Urieli, Marianne Vergez-Couret

Jochre, océrisation par apprentissage automatique : étude comparée sur le yiddish et l'occitan

Assaf Urieli Marianne Vergez-Couret[1]
CLLE-ERSS, UMR5263, Université de Toulouse 2 Le Mirail
assaf.urieli@univ-tlse2.fr, marianne.vergez@univ-tlse2.fr

RÉSUMÉ

Pour constituer des bases de textes pour des langues peu dotées comme le yiddish et l'occitan, il faut créer des outils et des ressources permettant une reconnaissance optique de caractères (océrisation) de qualité. Une des difficultés principales à surmonter pour ces deux langues est leur grande variation graphique (et dialectale pour l'occitan). Il est généralement admis qu'un lexique augmente la qualité de l'océrisation, mais la façon dont il faut tenir compte de la variation au sein du lexique n'est pas toujours clair. Dans cette étude, nous avons utilisé un logiciel d'apprentissage automatique supervisé, Jochre. Nous comparons plusieurs façons de constituer et d'utiliser les lexiques. La meilleure méthode nous permet d'atteindre une exactitude de 91,2 % (mots) et 97,4 % (lettres) sur le corpus yiddish, et de 93,2 % (mots) et 97,9 % (lettres) pour le corpus occitan.

ABSTRACT

Supervised machine learning OCR using Jochre : a comparative study of Yiddish and Occitan

To create textual databases for less-resourced languages such as Yiddish and Occitan, we need tools and resources enabling high-quality OCR (optical character recognition). One of the main difficulties to overcome for these two languages is their considerable spelling variation (and dialectal variation for Occitan). It is generally admitted that a lexicon can improve OCR quality, but it is not clear how to take such variation into account within the lexicon. In this study, we use Jochre, a supervised machine learning OCR system. We compare several methods of generating and using lexicons. The best method allows us to attain an accuracy of 91.2% (words) and 97.4% (letters) for the Yiddish corpus, and 93.2% (words) and 97.9% (letters) for the Occitan corpus.

MOTS-CLÉS : OCR, océrisation, yiddish, occitan, lexique, dialectes, apprentissage automatique.
KEYWORDS : OCR, Yiddish, Occitan, lexicon, dialects, supervised machine learning.

1 Introduction

Dans cette étude, nous nous intéressons à l'océrisation, à l'aide du logiciel Jochre, de deux langues peu dotées en ressources TAL : le yiddish et l'occitan, que nous présentons dans les sous-sections 1.1 et 1.2. Nous étudions plus particulièrement l'apport du lexique dans le contexte d'un système d'apprentissage automatique supervisé.

Il est souvent dit que l'océrisation des textes imprimés est un problème résolu, ce qui est vrai pour des documents récents en anglais, mais l'exactitude baisse énormément dès qu'il s'agit de langues peu dotées en ressources informatiques, ou de documents dont la qualité des

[1] Avec le soutien du Conseil Régional de Midi-Pyrénées et l'université de Toulouse Le Mirail dans le cadre d'un financement de post-doctorat.

images est dégradée. Cette exactitude peut se mesurer de deux façons : au niveau des mots et au niveau des lettres. Tanner, Muñoz et Ros (2009) mesurent une exactitude pour les lettres de 99,98 % pour des journaux anglais édités après 1950, de 95 % pour ceux édités entre 1900 et 1950, et de < 85 % pour ceux édités avant 1900. Les études récentes sont rares pour d'autres langues. On peut citer Boschetti et al. (2009) qui atteignent une exactitude pour les lettres oscillant entre 92 % et 95 % pour le grec ancien dans des éditions critiques historiques.

Quelques systèmes libres d'OCR ont vu le jour récemment, dont Tesseract (Smith, 2007), ainsi que OCROpus (Breuel, 2008) et Jochre[2], qui mettent l'accent sur des techniques d'apprentissage automatique supervisé. Pour cette étude, nous avons choisi Jochre car il présentait plusieurs avantages par rapport aux autres logiciels libres : il intégrait déjà un modèle linguistique pour le yiddish, un module d'intégration de lexiques, et un module d'évaluation sophistiquée. Ce logiciel sera présenté en section 2.

Les techniques d'apprentissage supervisé nécessitent la constitution de corpus annotés pour l'entraînement, et parallèlement de lexiques pour l'évaluation. La constitution de ces ressources est décrite dans la section 3, et les résultats de notre étude dans la section 4.

1.1 Occitan

1.1.1 La langue occitane

L'occitan est une langue romane, parlée dans le sud de la France, dans le Val d'Aran en Espagne et dans 12 vallées alpines d'Italie. Faute d'étude spécifique, il est difficile d'estimer le nombre de locuteurs occitanophones. Diverses études annoncent des chiffres allant de 600 mille à 2 millions de locuteurs (Martel, 2007). Cette langue connaît une variation interne relativement importante. Elle regroupe un ensemble de variétés organisées en dialectes :

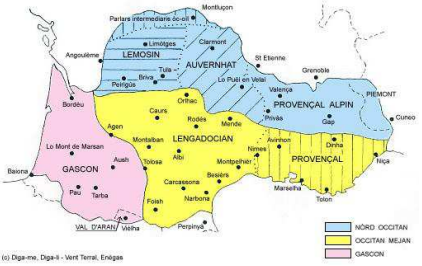

FIGURE 1 – Dialectes de l'occitan (Carles, 2005)

Dans cet article, nous nous focalisons sur deux dialectes de la langue occitane : le languedocien et le gascon dont voici quelques exemples de variation : lo filh/eth hilh ; luna/lua et cabra/craba (Bec, 1995).

L'occitan n'est pas une langue unifiée et standardisée. Mais elle est écrite depuis le moyen-âge et a produit une littérature très importante. L'occitan s'écrit alors dans une graphie dite "des troubadours". Cette graphie disparaît peu à peu que la production littéraire décroit. Au XIXème siècle, une première graphie normalisée voit le jour en Provence, la graphie dite

[2] https://github.com/urieli/jochre

mistralienne. Au XXème siècle apparaît une orthographe unifiée, dite graphie classique inspirée de la graphie des troubadours (Sibille, 2007). Dans le cadre de cette étude, nous avons souhaité utiliser uniquement des œuvres écrites ou transcrites en graphie classique pour nous focaliser sur la variation dialectale. Toutefois, nous notons tout de même des variations d'un auteur à un autre, par exemple, on trouve dans un œuvre gasconne en graphie classique des vestiges de la graphie dite béarnaise (tabé vs. tanben ; coundes vs. condes). Il demeure également des variations liées à la normalisation qui ne s'est pas faite en un jour, par exemple sur les conjugaisons (avian vs. avián) ou des variations de prononciation (contes vs. condes).

1.1.2 Projet BaTelÒc

Le projet BaTelÒc[3] (Bras, 2006 ; Bras et Thomas, 2011) vise la construction d'une base de textes en langue occitane (sur le modèle de Frantext) en rassemblant des œuvres écrites de tous genres (littérature, théâtre, contes, textes techniques et journalistiques…) des périodes modernes et contemporaines. Un million de mots a déjà été rassemblé à partir d'œuvres contemporaines existant au format numérique. Suffisamment de matières est disponible pour envisager de passer à plusieurs centaines de millions de mots. Mais une partie de cette matière n'existe pas encore au format numérique et une étape préalable de numérisation et d'océrisation est alors nécessaire. Nous ne connaissons pas à ce jour de logiciel d'océrisation capable de traiter convenablement la variation graphique et dialectale autrement qu'en rentrant manuellement des listes de mots. Nous avons donc souhaité entraîner un modèle pour l'occitan et tester l'impact de l'utilisation de lexiques de formes fléchies construits de façon hiérarchique pour la langue, les dialectes et les parlers[4] sur les performances de l'océrisation. Le TALÒc (Traitement automatique de la langue occitane) fait intervenir une dimension nouvelle par rapport au TAL des langues très dotées : la gestion de la variation dialectale et graphique.

1.2 Yiddish

1.2.1 La langue yiddish

Le yiddish est à l'origine la langue des communautés juives ashkénazes d'Europe centrale et de l'est. Au début du 20ème siècle, on estime le nombre de yiddishophones de 11 à 13 millions (Jacobs, 2005). Suite au génocide nazi, à la russification forcée de l'époque stalinienne, à la position anti-yiddish du mouvement sioniste, et à l'assimilation des juifs aux Etats-Unis, ce nombre s'est réduit de nos jours à quelques centaines de milliers, dont la plupart font partie des communautés « hassidiques », pourtant en pleine croissance démographique (Katz, 2004). Le yiddish est une langue hybride, avec un vocabulaire de base tiré de l'allemand médiéval, de l'hébreu biblique, de l'araméen, et des langues slaves des pays où vivaient les juifs, mais souvent classé comme langue germanique. Il s'écrit avec l'alphabet hébreu, mais avec certaines différences majeures : les voyelles, largement absentes dans l'hébreu en dehors des signes diacritiques, sont réintroduites dans le yiddish en tant que lettres à part entière, sauf pour les mots d'origine hébraïque, qui sont écrits comme dans les sources religieuses, sans les voyelles.

[3] Projet dirigé par Myriam Bras (CLLE-ERSS Université de Toulouse 2) depuis 2006.
[4] Le terme parler est utilisé pour désigner une variante d'un dialecte sur une aire géographique plus restreinte (un village, une vallée...).

A l'instar de l'occitan, le yiddish est une langue sans gouvernement ni académie. Il y a deux normes orthographiques officielles : celle du YIVO[5] et celle du gouvernement soviétique. Le YIVO a cherché à définir un yiddish moderne et standard qui engloberait tous les dialectes, alors que le gouvernement soviétique a cherché à effacer tout rappel de l'hébreu dans le yiddish afin de l'éloigner de toute connotation religieuse. Cependant, jusqu'à 1950, la plupart des éditeurs en dehors de l'Union Soviétique ont appliqué leurs propres normes orthographiques, parfois inspirées de l'allemand (redoublement des consonnes à l'intérieur des mots, ajout de la lettre hébraïque « hey » pour imiter un H muet), et rarement cohérentes. Au niveau dialectal, les éditeurs se voulaient presque tous « universels », et les différences dialectales, très présentes à l'oral, sont largement absentes à l'écrit.

1.2.2 Le Corpus du Yiddish Book Center

Le Yiddish Book Center à Amherst, Massachussetts a constitué un corpus à partir des livres en yiddish retrouvés partout dans le monde (Lansky, 2004). Ils ont amassé 1,5 million de livres à ce jour, dont une grande partie a été distribuée à plus de 600 bibliothèques, universitaires et autres. Les livres comprennent 18 000 titres uniques, dont 12 000 ont été scannés, et peuvent être téléchargés gratuitement en ligne[6]. Ces livres sont en grande majorité édités entre 1870 et 1960, période qui correspond à la renaissance de la culture et de la littérature yiddish. Notre corpus d'entraînement est tiré entièrement de ce corpus.

2 Le logiciel OCR Jochre

Le logiciel Jochre[7] (*Java Optical CHaracter REcogition*) est un logiciel OCR libre développé par Assaf Urieli, au départ pour océriser le corpus du Yiddish Book Center avec des techniques d'apprentissage automatique supervisé. Pour cette étude, nous l'avons adapté à l'occitan. L'analyse de Jochre s'effectue en trois étapes : 1) segmentation des images en paragraphes, lignes, groupes et formes ; 2) reconnaissance des lettres et 3) correction des mots à l'aide du lexique.

2.1 Segmentation

La segmentation utilise des techniques statistiques ad hoc adaptées à chaque tâche (détection de l'orientation, suppression des petites taches, …). A la sortie de la segmentation, on a une *forme* pour chaque lettre, et un *groupe* de formes pour chaque mot. Dans cet article, un *groupe* signifie donc « séquence de formes graphiques qui correspond à un seul mot dans la page ». Il arrive que Jochre ne soit pas capable de trouver la segmentation exacte d'une image. En particulier, deux lettres peuvent être fusionnées en une seule forme (trop d'encre lors de l'impression, tâche), et une lettre peut être scindée en deux formes (manque d'encre, livre abîmé). Les taux d'erreur de segmentation sont de 2,6 % pour le corpus yiddish, et de 12,5 % pour le corpus occitan. Cette étude ne cherche pas à améliorer la segmentation de Jochre : quand deux lettres sont fusionnées, Jochre va tenter de trouver les deux lettres à la fois, et quand une lettre est scindée, il va tenter de trouver les deux moitiés de la lettre

[5] YIVO : le Yidisher Visnshaftlekher Institut, ou Institut Scientifique Yiddish, établie à Vilnius en 1925 en tant que référence dans l'étude de la langue yiddish.

[6] http://www.yiddishbookcenter.org/books/search

[7] https://github.com/urieli/jochre

séparément lors de la reconnaissance des lettres.

2.2 Reconnaissance des lettres

La reconnaissance des lettres s'applique à chaque forme retrouvée lors de l'étape de segmentation. Elle utilise des techniques d'apprentissage automatique, et se divise donc en deux activités distinctes : l'entraînement et l'analyse.

Lors de l'étape d'*entraînement*, on sélectionne un corpus d'apprentissage pour la langue qu'on souhaite océriser en scannant des documents (quelques pages par document). Les images scannées sont ensuite chargées dans le logiciel web JochreTrain qui effectue automatiquement la segmentation de ces images (étape précédente). A travers l'interface JochreTrain, l'utilisateur attribue manuellement la bonne lettre à chaque forme. Ensuite, il faut choisir les *traits* qui caractérisent les formes. Un trait est n'importe quelle information concernant la forme qui peut aider le logiciel à choisir la bonne lettre. Par exemple, on peut imaginer un trait PetitPointEnHaut qui donne « vrai » si la forme contient un petit point en haut, et « faux » dans le cas contraire. Pour le français, un résultat « vrai » pour ce trait peut indiquer que la forme est un « i » à 95 % ou un « j » à 5 %. La dernière étape de l'entraînement est la construction du modèle statistique. Pour ceci, on utilise un *classifieur*, qui est le « moteur » chargé de décider de l'importance relative à accorder à chaque trait. Il va ainsi attribuer un poids à chaque trait et à chaque lettre, en cherchant à maximiser la vraisemblance de ces poids par rapport au corpus d'apprentissage. La grande matrice des poids par trait et par lettre est ce qu'on appelle le *modèle statistique*. Dans les expériences décrites ici, on utilise un classifieur par entropie maximale, basé sur Ratnaparkhi (1998).

Lors de l'étape d'*analyse*, l'utilisateur fournit à Jochre une nouvelle image à océriser. Jochre va segmenter l'image, appliquer les traits à chaque forme trouvée dans l'image segmentée, et appliquer le modèle statistique aux résultats des traits pour générer une distribution de probabilités des lettres pour cette forme (selon les probabilités observés dans le corpus).

2.2.1 Traits de base (baseline)

Pour la reconnaissance des lettres, nous avons utilisé une liste de traits de base, qui s'appliquent à n'importe quel alphabet et langue. Parmi ceux-ci se trouvent les traits *n-grammes*, qui indiquent les séquences de lettres les plus probables dans une langue donnée. Le *n* réfère au nombre des lettres dans la séquence. En français, par exemple, la séquence « pro » est bien plus probable que la séquence « pry ». Donc, si pendant l'analyse on a déjà choisi « p » pour la première forme et « r » pour la deuxième, le trait 3-grammes nous indique que la troisième lettre est plus probablement un « o » qu'un « y ».

La liste complète des traits de base est la suivante :

- Ngram2 : la lettre reconnue par Jochre pour la forme précédente
- Ngram3 : les lettres reconnues par Jochre pour les deux formes précédentes
- SectionRelativeBrightnessGrid : La noirceur relative de chaque section d'une grille (divisée en 9 sections verticales et 5 sections horizontales) par rapport à la noirceur de la section la plus sombre.
- VerticalSize : La hauteur verticale de la forme, divisée par l'œil typographique

(hauteur d'une lettre minuscule comme « x » sans jambage supérieur ni inférieur).

- VerticalElongation : Le rapport entre la hauteur et la largeur de la forme.
- BaselineDistance : La limite inférieure de la forme par rapport à la ligne de pied.
- LastShapeInSequence : Est-ce que cette forme est la dernière dans le groupe actuel ?
- TouchesBaseLine : Est-ce que cette forme touche la ligne de pied ?
- TouchesMeanLine : Est-ce que cette forme touche la ligne en haut de l'œil ?

2.2.2 Traits spécialisés pour l'alphabet hébreu

Pour le yiddish, nous avons aussi développé des traits spécifiques aux paires de lettres qui sont souvent confondues par le logiciel quand on applique uniquement les traits de base.

FIGURE 2 – Les lettres similaires *daled* et *reysh* dans l'alphabet hébreu

Dans la FIGURE 2 on voit deux lettres hébraïques très similaires, *daled* et *reysh*. On a donc développé un trait qui cherche une protubérance dans le coin en haut à droite de la lettre. Des traits semblables ont été développés pour distinguer 9 paires de lettres. Le développement de ces traits « spécialisés » est assez coûteux en temps. Du coup, pour l'occitan, nous avons basé notre évaluation uniquement sur les traits de base.

2.3 Correction à l'aide d'un lexique

Lors de la reconnaissance des lettres, Jochre applique une technique de « recherche par faisceau » (*beam search* ou *breadth-first search*)) (Bisiani, 1992). Pour chaque groupe, cette technique va fournir les n séquences de lettres les plus probables, construites à partir des distributions des probabilités des lettres pour chaque forme dont le groupe est constitué. Il s'agit maintenant de choisir le bon mot parmi ces séquences. En particulier, Jochre peut utiliser un lexique pour modifier les scores de chacune des n séquences. Un *lexique* ici est une liste de formes lexicales fléchies possibles dans une langue donnée, avec éventuellement une fréquence associée à chaque forme. Si la séquence est inconnue dans le lexique, on va multiplier son score par un coefficient de réduction < 1.

acordat	**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 %

TABLE 1 – Analyse avec un coefficient de réduction de 0,5 et un faisceau de 3

Le TABLE 1 ci-dessus montre l'analyse du mot « acordat » par Jochre, avec un lexique qui contient ce mot, un coefficient de réduction des mots inconnus de 0,5, et une largeur de faisceau de 3. A la sortie de la reconnaissance des lettres, Jochre propose donc les 3 séquences de lettres les plus probables : « acordot » (72,0 %), « acordat » (70,1 %), et

« acordet » (64,3 %). Noter que ces scores se calculent en prenant la moyenne harmonique des probabilités pour chaque lettre de la séquence. Sans lexique, c'est « acordot » qui aurait été choisi. Avec le lexique, les scores pour les mots inconnus sont multipliés par un coefficient de 0,5, et c'est « acordat » qui se trouve en tête. Il est aussi possible de privilégier les mots fréquents, si le lexique indique la fréquence relative des mots. Dans ce cas, Jochre nous permet d'utiliser la formule : $\text{Score}_{\text{final}} = \text{Score}_{\text{initial}} \times (1 + \log_k(\text{freq}))$

où $Score_{initial}$ est le score fourni par la fonction de reconnaissance des lettres pour la séquence en question, k la base du logarithme, fourni comme paramètre à Jochre, et $freq$ la fréquence du mot indiqué par le lexique. Plus k est élevée, moins la fréquence va compter dans le score final. Par exemple, pour k=10, le score d'un mot dont la fréquence est 1 restera inchangé, le score d'un mot dont la fréquence est 10 sera multiplié par 2, et le score d'un mot dont la fréquence est 100 sera multiplié par 3.

3 Corpus annoté et lexiques

3.1 Préparation des corpus

Les corpus d'entraînement et d'évaluation sont constitués à partir 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). Pour l'occitan, le corpus est également diversifié du point de vue dialectal : 39 pages pour le gascon et 41 pages pour le languedocien.

	Corpus Yiddish	**Corpus Occitan**
Nombre de livres numérisés	14	10[8]
Années d'édition	1910-1955	1960-2000
Lieu d'édition	Europe, Amérique du Nord, Amérique du Sud	France
Nombre de pages	95	80
Nombre de mots	27 400	20 400
Nombre de lettres	123 400	85 500

TABLE 2 – Caractéristiques des corpus

Les corpus ont été corrigés après une première analyse automatique du corpus en les confrontant là où l'annotation manuelle et l'annotation automatique diffèrent : l'annotation automatique permet dans ce cas de corriger des erreurs humaines et d'améliorer la qualité du corpus d'entraînement.

3.2 Préparation des lexiques de formes fléchies

3.2.1 Yiddish

Pour le Yiddish, nous avons utilisé les entrées en yiddish du dictionnaire bilingue yiddish-français de Yitskhok Niborski et Bernard Vaisbrot (Niborski et Vaisbrot, 2002). Ce

[8] 4 auteurs présents dans le corpus, Rouquette, Blader, Laux et Mouly, ont un statut spécial car nous disposons pour ces auteurs de lexiques extraits de leurs œuvres, cf. section 3.2.2.

dictionnaire, avec 37 000 mots, a la couverture lexicale la plus large de tous les dictionnaires yiddish bilingues parus jusqu'à présent. La version informatique de la partie yiddish du dictionnaire nous a été fournie par Harry Bochner, éditeur de la traduction en anglais de ce dictionnaire (Beinfeld et Bochner, 2013). Nous avons développé un logiciel pour transformer cette liste automatiquement en une liste de 396 500 formes fléchies. Le dictionnaire de Niborski et Vaisbrot contient très peu de noms propres. Pour combler ce manque, nous avons compilé, avec l'aide du Yiddish Book Center, des listes de prénoms et diminutifs tirées au départ de Harkavy (1928), des noms de famille et des noms géographiques.

3.2.2 Occitan

L'occitan est encore à ce jour une langue peu dotée en matière de ressources lexicales et informatisées même s'il existe de nombreux dictionnaires sous format papier (Bras et Thomas, 2007). Pour commencer, nous avons tiré profit des textes déjà présents dans BaTelÒc pour construire des lexiques de formes fléchies pour le languedocien et le gascon en considérant toutes les formes trouvées dans ces œuvres.

	Languedocien	Gascon
Total (sans les doublons)	70400 formes	10100 formes
Dont les auteurs[9]	Laux (7700 formes)	Blader (5 300 formes)
	Mouly (9600 formes)	
	Rouquette (11 000 formes)	

TABLE 3 - Lexiques de formes fléchies par dialecte issus de BaTelÒc

Nous avons ensuite constitué des lexiques à partir des entrées de plusieurs dictionnaires numérisés : pour le gascon, le dictionnaire français/occitan du gascon toulousain (Rei Bèthvéder, 2004) qui contient des noms communs et des noms propres et pour le languedocien, le dictionnaire français/occitan (Laux, 2005) et le lexique occitan/français des mots occitans de Max Rouquette (Rouquette, 2010). Lorsque les informations de flexion étaient signalées, nous avons automatiquement construit les féminins des noms et des adjectifs. Enfin, nous avons construit les pluriels pour tous les noms et les adjectifs.

	Languedocien	Gascon
Dictionnaire/Lexique	Dictionnaire Laux (81 200 formes)	Dictionnaire Gascon Toulousain (21 900 formes)
	Lexique de G. Rouquette (3 800 formes)	
Total (sans les doublons)	84 376 formes	21 900 formes

TABLE 4 - Lexiques de formes fléchies par dialecte issus de dictionnaires et lexiques

Tous les lexiques ainsi constitués contiennent des informations de fréquence : la fréquence absolue pour les formes fléchies extraites des textes de BaTelÒc et une fréquence de 1 pour les entrées et leur(s) flexion(s) des dictionnaires ou lexiques bilingues.

[9] Uniquement les auteurs qui sont également présents dans le corpus.

Enfin, nous avons fusionné les lexiques de la façon suivante afin de tester pour les deux sous-corpus gascon et languedocien l'apport de chaque :

- Lex_Global (150 700 formes) : tous les lexiques réunis (tous dialectes confondus).
- Lex_Languedocien (135 300 formes): tous les lexiques en dialecte languedocien.
- Lex_Gascon (28 900 formes) : tous les lexiques en dialecte gascon.

Nous avons également fusionné des lexiques par auteur afin de tester l'impact de la précision de ces lexiques (représentant précisément le parler et la graphie de l'auteur) pour chaque sous-corpus correspondant à l'auteur.

- Lex_Rouquette (17 700 formes) : les lexiques de formes fléchies des textes de BaTelÒc et les formes issues du lexique des mots occitans de Rouquette.
- Lex_Laux (84 800 formes): les lexiques de formes fléchies des textes de BaTelÒc et les formes fléchies issues du dictionnaire du même auteur.
- Lex_Molin (9 600 formes): les lexiques de formes fléchies des textes de BaTelÒc.
- Lex_Blader (5 300 formes) : les lexiques de formes fléchies des textes de BaTelÒc.

4 Résultats

Pour chaque corpus, nous avons effectué une validation croisée en le divisant en six parts égales : à chaque fois nous avons entraîné sur 5/6èmes et évalué sur le 1/6ème restant, chaque 1/6ème étant tour à tour évalué. Les scores indiqués ici sont les scores totaux des 6 validations (nombre d'erreurs/nombre total). Ce score total est toujours très proche du score moyen des 6 validations avec un écart type inférieur à 1 %.

4.1 Yiddish

Nous avons mesuré les taux d'erreur 1) avec et sans lexique, et 2) avec et sans traits spécialisés pour le corpus entier et pour chaque œuvre séparément. Les meilleurs résultats sont de 91,20 % pour les mots et 97,44 % pour les lettres. Si on prend la mesure sans lexique et sans traits spécialisés comme baseline, ceci représente un gain de 31,7 % (mots) et de 30,6 % (lettres) par rapport à toutes les erreurs qu'il restait à corriger. Il y a une variation conséquente d'exactitude par œuvre, allant, pour les mots, de 78,93 % pour un livre de poésie alternant une police italique et une police décorative, à 97,78 % pour un livre relativement récent et bien conservé dans une police standard de type « Times ».

	Traits de base		Traits spécialisés	
	Mots	Lettres	Mots	Lettres
Sans lexique	87,11	96,31	89,74	97,09
Avec lexique	89,14	96.84	**91,20**	**97,44**

TABLE 5 - Taux de réussite par méthode pour le yiddish

Comme indiqué dans la TABLE 5, les traits spécialisés apportent un peu plus que le lexique, mais les meilleurs résultats sont obtenus par la combinaison des deux. Les traits spécialisés

permettent de mieux identifier les lettres au départ, alors que le lexique ne peut s'appliquer qu'aux *n* analyses les plus probables déjà identifiées.

4.2 Occitan

Pour l'occitan, nous avons mesuré les taux d'erreur avec et sans lexique. Les meilleurs résultats pour la totalité du corpus sont de 93,15 % pour les mots et 97,93 % pour les lettres. On note, comme pour le yiddish, une variation par œuvre allant, pour les mots, de 88,21 % pour une œuvre en police italique à 97,21 % pour une œuvre en police standard. Si on prend la mesure sans lexique comme baseline, ceci représente un gain de 19 % (mots) et de 16,2 % (lettres) par rapport à toutes les erreurs qu'il restait à corriger.

	Ensemble du corpus	
	Mots	**Lettres**
Sans lexique	91,54	97,53
Lex_Gascon	92,72	97,81
Lex_Languedocien	92,83	97,86
Lex_Global	**93,13**	**97,93**

TABLE 6 – Taux de réussite pour le corpus occitan total

Les résultats montrent un apport systématique des lexiques pour l'occitan, que ce soit un lexique global ou un lexique par dialecte. Les meilleurs résultats pour l'ensemble du corpus sont obtenus avec le lexique englobant languedocien et gascon. Nous avons, dans un second temps, comparé pour chaque sous-corpus gascon et languedocien l'impact des lexiques languedocien et gascon.

	Corpus Languedocien		Corpus Gascon	
	Mots	**Lettres**	**Mots**	**Lettres**
Sans lexique	92,08	97,64	90,99	97,41
Lex_Gascon	93,07	97,85	**92,36**	**97,78**
Lex_Languedocien	**94,10**	**98,15**	91,53	97,56
Lex_Global	94,08	98,13	92,16	97,71

TABLE 7 – Taux de réussite pour le corpus occitan par dialecte

Pour le sous-corpus languedocien, les meilleurs résultats sont de 94,10 % pour les mots et de 98,15 % pour les lettres (gain de 25,5 % (mots) et de 21,6 % (lettres) par rapport à toutes les erreurs qu'il restait à corriger) avec le lexique en languedocien mais ces résultats sont tout juste sensiblement meilleurs à ceux obtenus avec le lexique global (qui de fait est majoritairement composé de formes en languedocien dû au déséquilibre de la présence des deux dialectes dans le lexique). Pour le sous-corpus gascon, les meilleurs résultats sont de 92,36 % pour les mots et 97,78 % pour les lettres (gain de 15,2 % (mots) et 14,2 % (lettres) par rapport à toutes les erreurs qu'il restait à corriger) avec le lexique gascon. L'utilisation d'un lexique permet toujours d'apporter un gain, même en utilisant le lexique gascon sur le

corpus languedocien et le lexique languedocien sur le corpus gascon. En revanche, le lexique languedocien permet un gain de 25,5 % (mots) sur le corpus languedocien contre 6 % (mots) sur le corpus gascon. Inversement, le lexique gascon permet un gain de 15,2% (mots) sur le corpus gascon contre 12,5 % (mots) sur le corpus languedocien. Cela encourage prioritairement la création de lexiques de taille importante et en second lieu de lexiques par dialecte.

Enfin, nous regardons spécifiquement les sous-corpus des auteurs Blader (pour le gascon), Molin, Laux et Rouquette (pour le languedocien). Pour chacun des sous-corpus, nous cherchons le lexique qui a permis d'obtenir les meilleurs résultats pour les mots.

	Corpus_Blader	**Corpus_Laux**	**Corpus_Molin**	**Corpus_Rouquette**
Lex_Global	**97,62**	94,52	96,05	89,56
Lex_Languedocien	96,75	94,25	**96,1**	89,67
Lex_Gascon	97,53	93,89	95,44	87,11
Lexique de l'auteur	97,39	**94,34**	95,95	**89,95**

TABLE 9 – Pourcentage de réussite (mots) pour le corpus occitan par auteur

De bons résultats sont obtenus en utilisant le lexique global pour le sous-corpus en gascon (Corpus_Blader) et le lexique languedocien pour les trois autres sous-corpus. L'utilisation de petit lexique spécialisé par auteur permet une augmentation sensible des résultats pour le Corpus_Laux et le Corpus_Rouquette bien qu'ils soient jusqu'à 10 fois plus petits. Ces résultats encouragent, lorsque c'est possible, l'utilisation d'un lexique spécifique à l'auteur qui grossit à mesure que son œuvre est numérisée.

4.3 Autres paramètres

Concernant le coefficient de réduction du score pour les mots inconnus (*cf.* section 2.3.), nous avons testé des coefficients variants de 0,1 à 1,0, où un coefficient de 1,0 met les mots connus et inconnus au même niveau (comme s'il n'y avait pas de lexique). La FIGURE 3 montre l'exactitude des mots (connus, inconnus et les deux confondus) selon ce coefficient. Plus le coefficient augmente, plus l'exactitude des mots connus baisse, et celle des mots inconnus augmente. Comme il y a 27,0 % de mots inconnus dans le corpus yiddish et 14,9 % dans le corpus occitan, l'apport des mots connus est plus important. L'équilibre total est atteint dans les deux cas avec un coefficient de 0,75, ce qui est assez surprenant, vu la différence significative dans le mode de construction des deux lexiques. Ce qui est surtout étonnant dans ce graphique, c'est la différence importante qui subsiste entre les mots connus et les mots inconnus dans les deux corpus quand on atteint un coefficient de 1,0. Il faudrait analyser les données pour émettre des hypothèses quant à cette différence[10].

Nous n'avons pas testé l'apport de la fréquence pour le yiddish, car nos lexiques pour le yiddish sont tous construit à partir de dictionnaires. Pour l'occitan, prendre la fréquence en compte fait baisser les scores systématiquement. Plus on accord de l'importance à la fréquence, plus le score baisse. Même en utilisant un logarithme à base 1000 (c'est-à-dire

[10] Nous remercions le relecteur anonyme pour son hypothèse que nous pourrons tester ultérieurement, à savoir que les mots inconnus sont peut-être des mots plus longs.

que si un mot avait une fréquence de 1000, son score serait deux fois plus élevé que s'il avait une fréquence de 1), l'exactitude baisse de presque 1 %.

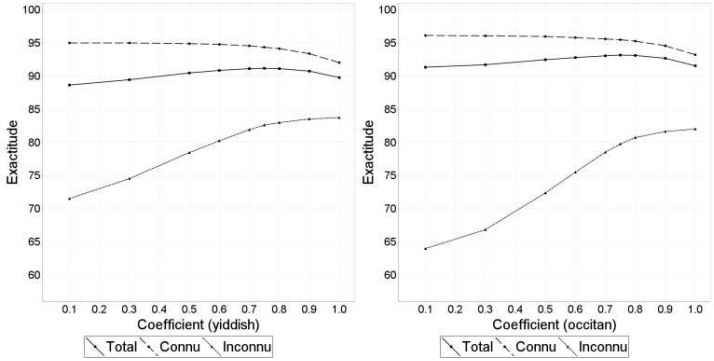

FIGURE 3 - Exactitude des mots selon le coefficient de réduction des mots inconnus

Concernant la recherche par faisceau, nous avons testé des faisceaux de 1, 2, 5, 10, 20 et 50. Un faisceau de 10 donne des résultats supérieurs aux autres faisceaux testés, ce qui voudrait dire qu'avec des faisceaux plus larges, le lexique choisit de façon erronée des mots peu probables, alors qu'avec des faisceaux moins larges, le lexique n'a pas toujours accès au bon mot dans la liste des séquences qui lui est fournie. Pour connaître l'apport maximal possible pour un lexique, il faut savoir, parmi les erreurs, combien de faisceaux contiennent la bonne réponse. Pour les erreurs en yiddish, avant d'appliquer le lexique, le faisceau de 2 contient 40 % des bonnes réponses, 5 contient 59 %, 10 contient 65 %, 20 contient 68 %, et 50 contient 69 %. Après l'application du lexique actuel au faisceau de 10, parmi les erreurs qui reste, le faisceau contenait la bonne réponse dans 59 % des cas, ce qui correspond à l'apport maximal possible pour un lexique à plus grande couverture. Pour l'occitan avec un faisceau de 10, ce pourcentage passe de 69 % sans lexique à 62 % avec le lexique actuel.

Le dernier point important à noter est la différence entre les mots bien segmentés et ceux qui contiennent au moins une lettre mal segmentée. Dans le corpus yiddish, les mots mal segmentés ne représentent que 2,6 % du corpus, mais leur taux d'erreur est de 58,7 % contre 7,5 % pour les autres. Dans le corpus occitan, les mots mal segmentés représentent 12,5 % du corpus, avec un taux d'erreur de 26,8 % contre 4,0 % pour les autres.

5 Conclusion et perspectives

Dans cette étude, nous nous sommes intéressés à l'océrisation, à l'aide du logiciel Jochre, de deux langues peu dotées en ressources TAL : le yiddish et l'occitan. Pour le yiddish, nous avons confirmé l'apport du lexique et des traits spécifiques à l'alphabet hébreu, mais la question reste ouverte sur la couverture plus faible du lexique comparé à l'occitan, alors que ce lexique est tiré d'un dictionnaire à large couverture, augmenté de listes de noms propres. Pour l'occitan, nous souhaitions évaluer l'apport des lexiques en faisant jouer d'une part les dialectes et d'autre part les parlers de plusieurs auteurs pour lesquels nous disposions de lexiques et d'œuvres numérisées. Les résultats montrent en premier lieu l'importance de disposer de lexiques de taille importante. Néanmoins, certains des résultats tendent à indiquer l'importance de la précision des lexiques, au niveau du dialecte, mais également au

niveau du parler spécifique de l'auteur. Il sera intéressant de regarder qualitativement là où le lexique apporte un gain ou induit le système en erreur (ce qui est le risque majeur de l'utilisation d'un lexique non approprié au dialecte ou au parler). De plus, seule cette analyse qualitative permettra de mesurer l'influence de la variation graphique *vs.* dialectale sur les performances de l'OCR.

Pour les deux langues, nos résultats mettent en avant deux classes de mots présentant une exactitude nettement plus basse : les mots mal segmentés et les mots inconnus dans le lexique. Pour les mots mal segmentés, outre améliorer les algorithmes ad hoc de segmentation, on pourrait imaginer un système d'apprentissage de la segmentation, entraîné sur le corpus d'apprentissage déjà constitué. Ce système tenterait de reconnaître des formes à scinder ou à fusionner, en étudiant les traits graphiques des formes déjà annotées en tant que lettre fusionnée ou scindée. Pour les mots inconnus, il faudrait procéder à une typologie, afin de cerner l'effort lexical qui serait le plus utile pour ajouter ce type de mot au lexique, sachant qu'il reste encore 59 % d'erreurs pour le yiddish et 62 % pour l'occitan qui peuvent être corrigées par un lexique à plus large couverture. Il faudrait également expliquer pourquoi une différence importante subsiste entre mots connus et inconnus même quand le lexique n'est pas utilisé : y-a-t-il un type de mot inconnu qui est intrinsèquement difficile à océriser et si oui, pourquoi ?

Au niveau du développement informatique, on a pu constater l'apport majeur des traits spécialisés à l'océrisation de l'alphabet hébreu. Il serait intéressant de développer des traits spécialisés pour l'alphabet latin, dans l'espoir de voir un apport semblable pour l'occitan.

Finalement, il nous semble intéressant d'explorer ce que l'océrisation peut apporter au lexicographe, dans l'optique d'enrichir les dictionnaires existants. Pendant l'analyse d'une œuvre, Jochre peut sortir une liste des formes inconnues dans le lexique, dont des erreurs d'analyse, des noms propres, des graphies inhabituelles, et des formes qui manquent dans les dictionnaires actuels. Ces dernières sont particulièrement intéressantes pour le lexicographe, d'autant plus qu'elles se trouvent dans le contexte d'une phrase tirée souvent d'une œuvre littéraire.

Remerciements

Pour le Yiddish, nous tenons à remercier Aaron Lansky, Catherine Madsen, Katie Palmer Finn, Joshua Price, Agnieszka Ilwicka du Yiddish Book Center pour leur aide dans la préparation du corpus et du lexique ; Yitskhok Niborski (INALCO) et Gilles Rozier (Bibliothèque Medem) pour leur accord concernant l'utilisation du dictionnaire de Niborski et Vaisbrot ; Harry Bochner pour avoir fourni ce dictionnaire en format XML ; et Paul Glasser du YIVO pour avoir fourni une liste des noms géographiques.

Pour l'occitan, nous tenons à remercier N. Rei Bèthvéder, C. Laux et G. Rouquette, l'IDECO et l'IEO du Tarn qui ont mis à disposition des versions électroniques d'œuvres, de dictionnaires et de lexiques pour la recherche et Myriam Bras pour les textes du projet BaTelÒc.

Références

BEC, P. (1995). *La langue occitane*. Que sais-je n°1059. Paris.

BEINFELD, S., et BOCHNER, H. (Eds.) (2013). *Comprehensive Yiddish-English Dictionary*. Indiana University Press.

BISIANI, R. (1992). Beam search. *Encyclopedia of Artificial Intelligence*, pages 1467-1468. Wiley-Interscience, 2nd edition. Editor: S. C. Shapiro.

BOSCHETTI, F., ROMANELLO, M., BABEU, A., BAMMAN, D., ET CRANE, G. (2009). Improving OCR accuracy for classical critical editions. *In Research and Advanced Technology for Digital Libraries*, pp. 156-167. Springer Berlin Heidelberg.

BRAS, M. (2006). Le projet TELOC : construction d'une base textuelle occitane. *In Langues et Cité : bulletin de l'observation des pratiques linguistiques,* 8, p9.

BRAS, M. et THOMAS, J. (2007). Diccionaris, corpora, e basas de donadas textualas, *In Linguistica Occitana, 5,* p. 1-22.

BRAS, M. et THOMAS, J. (2011). Batelòc : cap a una basa informatisada de tèxtes occitans. *In* A. Rieger (ed.) *L'Occitanie invitée de l'Euregio. Liège 1981 - Aix-la-Chapelle 2008 Bilan et perspectives*, Actes du IXème Congrès International de l'AIEO, Aache, Shaker.

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

CARLES, S. (2005). *Diga-me, diga-li*, Vent Terral.

HARKAVY, A. (1928). *Yiddish-English-Hebrew Dictionary.* Hebrew Publishing Co., New York.

JACOBS, N. (2005) *Yiddish: a Linguistic Introduction*, Cambridge University Press.

KATZ, D. (2004). *Words on Fire: The Unfinished Story of Yiddish* , New York: Basic Books.

LANSKY, A. (2004). *Outwitting History: The Amazing Adventures of a Man Who Rescued a Million Yiddish Books.* Chapel Hill: Algonquin Books of Chapel Hill.

LAUX, C. (2005). *Dictionnaire Français-Occitan*. IEO del Tarn.

MARTEL, P. (2007). L'occitan, qui parle ? *In Langues et Cité : bulletin de l'observation des pratiques linguistiques, 10,* p3.

NIBORSKI, Y. et VAISBROT B. (2002). *Dictionnaire yiddish–français.* Bibliothèque Medem, Paris.

RATNAPARKHI, A. (1998) *Maximum entropy models for natural language ambiguity resolution*, PhD Thesis, University of Pennsylvania.

REI BETHVEDER, N. (2004). *Dictionnaire Français/Occitan Gascon Toulousain*. IEO edicions.

ROUQUETTE, J.-G. (2010). *Lexique Occitan-Français de Max Rouquette*. Association Amistats Max Rouquette.

SIBILLE, J. (2007). L'occitan, qu'es aquò ? In *Langues et Cité : bulletin de l'observation des pratiques linguistiques, 10*, p2.

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

TANNER, S., MUÑOZ, T., & ROS, P. H. (2009). Measuring mass text digitization quality and usefulness. *D-Lib Magazine*, *15*(7/8), 1082-9873.

