Outiller l'occitan : nouvelles ressources et lemmatisation

Aleksandra Miletić

Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

Outiller l'occitan : nouvelles ressources et lemmatisation

Aleksandra Miletić
Department of Digital Humanities, University of Helsinki
Yliopistonkatu 3, 00014 Helsinki, Finland
aleksandra.miletic@helsinki.fi

RÉSUMÉ

Ce travail présente des contributions récentes à l'effort de doter l'occitan de ressources et outils pour le TAL. Plusieurs ressources existantes ont été modifiées ou adaptées, notamment un tokéniseur à base de règles, un lexique morphosyntaxique et un corpus arboré. Ces ressources ont été utilisées pour entraîner et évaluer des modèles neuronaux pour la lemmatisation. Dans le cadre de ces expériences, un corpus plus large (2 millions de tokens) provenant du Wikipédia a été annoté en parties du discours et en lemmes. Le corpus est accessible sur Zenodo.

ABSTRACT

New Resources and Lemmatization Experiments for Occitan

This paper presents recent contributions to the creation of NLP tools and resources for Occitan. Several existing resources were modified or adapted, in particular a rule-based tokenizer, a morphosyntactic lexicon and a treebank. These resources were used to train and evaluate neural lemmatization models. As part of these experiments, a large corpus based on Wikipedia (2 million tokens) was POS-tagged and lemmatized. This new resource is shared through Zenodo.

MOTS-CLÉS : langues peu dotées, occitan, lemmatisation.

KEYWORDS: low-resourced languages, Occitan, lemmatization.

1 Introduction

L'occitan est une langue romane parlée majoritairement dans le sud de la France ainsi qu'en Italie et en Espagne. Il n'est pas standardisé et connaît une variation interne importante. L'occitan possède une riche littérature depuis les troubadours médiévaux, il est enseigné, de l'école à l'Université, mais n'a pas de statut officiel en France. Comme c'est souvent le cas des langues dans cette situation, les efforts pour construire des ressources et outils du TAL ont démarré tardivement par rapport à la majorité des langues officielles d'Europe. Néanmoins, à travers deux projets récents, RESTAURE [1] et LINGUATEC [2], les premiers corpus annotés (Bernhard *et al.*, 2018; Miletic *et al.*, 2020b) et ressources lexicales (Vergez-Couret, 2016; Bras *et al.*, 2020) ont été réalisés, ce qui a permis l'entraînement des premiers modèles d'étiquetage en parties du discours (Vergez-Couret & Urieli, 2015) et d'analyse syntaxique en dépendances (Miletic *et al.*, 2019b, 2020b). Notons cependant que ces expériences

1. RESTAURE : RESsources informatisées et Traitement AUtomatique pour les langues REgionales, convention ANR-14-CE24-0003-01. http://restaure.unistra.fr/

2. LINGUATEC : projet européen EFA 227/16 Développement de la coopération transfrontalière et du transfert de connaissances en technologies du langage. https://linguatec-poctefa.eu/fr/projet/

Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : travaux de recherche originaux – articles longs, pages 217–231.

s'appuient sur des méthodes d'apprentissage statistiques. Par ailleurs, à notre connaissance il n'y a pas encore de travaux publiés sur la lemmatisation de l'occitan.

La lemmatisation consiste à trouver la forme de base d'une forme fléchie donnée. La définition de la forme de base pour une catégorie grammaticale donnée peut varier d'une langue à l'autre. Elle peut consister, par exemple, à identifier le masculin singulier d'un adjectif (*bèlas* 'belles' > *bèu* 'beau') ou à retrouver l'infinitif d'une forme verbale (*calèva* 'fallait' > *caler* 'falloir'). Cette forme de traitement est particulièrement importante pour les langues à morphologie riche : elle permet de regrouper toutes les formes de surface provenant du même lemme et de diminuer ainsi la dispersion des données.

Les outils de lemmatisation basés sur l'apprentissage automatique statistique peuvent se diviser en deux groupes principaux : approches par arbres d'édition (*edit trees*) et approches par transduction de chaînes de caractères. La première méthode (Gesmundo & Samardzic, 2012; Grzegorz Chrupala & van Genabith, 2008; Müller *et al.*, 2015) consiste à identifier la série d'opérations d'édition qui permet d'obtenir le lemme à partir de la forme fléchie. Cette série d'opérations est attribuée à la paire *(forme fléchie, lemme)* en tant qu'étiquette. L'outil apprend l'arbre d'édition qui s'applique à chaque paire, ce qui permet d'aborder la lemmatisation comme une tâche de classification. Avec les méthodes neuronales, la lemmatisation a été redéfinie comme une tâche de transduction de chaînes de caractères (Bergmanis & Goldwater, 2018; Manjavacas *et al.*, 2019). L'avantage principal de ces méthodes par rapport à la génération précédente réside dans une meilleure capacité de généralisation aux tokens inconnus. Ce fait est particulièrement pertinent pour les langues à variation interne forte comme l'occitan, où on s'attend à un taux élevé de tokens inconnus.

Dans ce travail, nous décrivons des expériences en lemmatisation à l'aide d'outils basés sur des méthodes d'apprentissage neuronales. Nous détaillons également plusieurs modifications apportées aux outils et ressources existantes dans le cadre de ces expériences. Enfin, nous décrivons un corpus de 2 millions de tokens extrait du Wikipédia, que nous avons étiqueté, lemmatisé et diffusé.

2 Occitan : langue régionale peu dotée

L'occitan est une langue romane qui appartient au groupe gallo-roman, parlée dans une aire linguistique incluant 32 départements de la moitié sud de la France (à l'exception des zones catalane et basque), dans plusieurs vallées alpines d'Italie et au Val d'Aran en Espagne (cf. figure 2, empruntée à (Bernhard *et al.*, 2021)). La langue connaît une riche variation diatopique couplée à l'existence de plusieurs graphies. Aucune de ses variétés n'est reconnue comme standard, même si des standards dialectaux tendent à émerger. Ces facteurs contribuent à un important degré de variation de formes de surface, ce qui rend la création des ressources et outils pour le TAL plus exigeante que pour les langues standardisées.

2.1 Propriétés linguistiques et axes de variation

L'occitan est une langue *pro-drop* : la réalisation du sujet dans la phrase n'est pas obligatoire. Les formes verbales finies portent des marques de temps, de personne et de nombre. Dans de nombreux dialectes, le nombre et le genre sont marqués sur tous les éléments du groupe nominal. À la différence du français, l'occitan préserve l'utilisation du passé simple et du subjonctif de l'imparfait, y compris à l'oral. L'exemple en provençal donné en figure 1 illustre certaines de ces propriétés.

(1) a. Totei lei personas naisson liuras e egalas en dignitat e en drech.
b. Tóuti li persouno naisson liéuro e egalo en dignita e en dre.
‘Toutes les personnes naissent libres et égales en dignité et en droit.’

FIGURE 1 – Première phrase de la Déclaration des droits de l’homme en provençal, en graphie classique (a) et en graphie mistralienne (b)

La classification dialectale la plus répandue (Bec, 1995) regroupe les variétés de l’occitan en six principaux dialectes : l’auvergnat, le gascon, le languedocien, le limousin, le provençal et le vivaro-alpin (v. figure 2) . Ces dialectes connaissent à leur tour une variation interne plus ou moins prononcée.

FIGURE 2 – Carte de dialectes

La variation diatopique est présente à tous les niveaux de fonctionnement linguistique. Au niveau lexical, on trouve des cas de figure comme celui du mot *pomme de terre* : en gascon, on utilise la forme *mandòrra*, alors qu’en languedocien on utilise plutôt *trufa/trufet* ou *patana/patanon*. Différents dialectes ayant subi des processus phonétiques différents, on retrouve des séries de formes spécifiques à chaque dialecte, à l’instar du mot *fils*, qui correspond à *hilh* en gascon, à *filh* en languedocien et en limousin, et à *fiu* en provençal. Au niveau morphosyntaxique, la flexion verbale varie au niveau inter-dialectal, mais aussi intra-dialectal. À partir du paradigme flexionnel le plus répandu dans chaque dialecte, la forme de la première personne du pluriel du présent du verbe *être* correspond à *èm* en gascon, *sem* en limousin, *sèm* ou *sièm* en languedocien et *siam* en provençal. La variation syntaxique la plus prononcée est celle du gascon par rapport aux autres dialectes. À titre d’illustration, les clitiques objet et les pronoms réflexifs se trouvent en position post-verbale plus souvent qu’en languedocien, où ils sont typiquement préverbaux (cf. gasc. *ne cau pas està**’s** darrèr mieidia* vs lang. ***vos*** *cal pas demorar après miègjorn*, ‘vous ne devriez pas rester après midi’). Le gascon, à la différence de tous les autres dialectes de l’occitan, exhibe également des particules énonciatives, qui marquent la modalité de la phrase. Ces particules occupent la position entre le sujet (s’il est réalisé) et le verbe, et elles sont obligatoires dans certains parlers gascons. Comparez, *e.g.*, gasc. *Lo vent* ***que*** *s’èra lhevat et* ***que*** *hasó drin fresc vs* lang. *Lo vent s’èra levat et fasiá un pauc freg* ‘Le vent s’était levé et il faisait un peu froid’.

En outre, plusieurs normes d’écriture existent, dont deux sont dominantes aujourd’hui : la norme dite *mistralienne*, proche des conventions orthographiques françaises, et la norme dite *classique*, fondée sur l’orthographe utilisée par les troubadours occitans (Sibille, 2002). L’exemple 1 illustre la

différence entre ces deux normes en reprenant la même phrase dans le même dialecte (le provençal) en graphie classique (1a) et en graphie mistralienne (1b). À ces différences s'ajoute la variation au niveau individuel : il n'est pas rare qu'un rédacteur adopte une graphie qui lui est propre.

2.2 Situation de l'occitan en TAL

Bien que l'occitan ait été reconnu comme appartenant à l'héritage culturel de la France [3], il n'est pas reconnu comme langue officielle dans ce pays, où réside la majorité de ses locuteurs. Comme il est souvent le cas dans cette situation, l'outillage de l'occitan a démarré plus tardivement par rapport à celui de la majorité de langues officielles européennes. Deux projets majeurs ont permis de constituer, par exemple, un corpus étiqueté en parties du discours (Bernhard *et al.*, 2018), un lexique (Vergez-Couret, 2016; Bras *et al.*, 2020) et un corpus arboré (Miletic *et al.*, 2020a,b). Des expériences en étiquetage en parties du discours (Vergez-Couret & Urieli, 2015) et en analyse syntaxique automatique (Miletic *et al.*, 2019b, 2020b) ont donné lieu aux premiers modèles de traitement automatique de l'occitan. Un modèle de synthèse de la parole a été développé récemment (Corral *et al.*, 2020).

En ce qui concerne les corpus larges, l'occitan dispose d'une base de données textuelles intérrogeable en ligne (Bras & Vergez-Couret, 2016). Un corpus de 2 millions de tokens extrait des articles du Wikipédia en occitan est diffusé au sein de *Leipzig Corpora Collection* (Goldhahn *et al.*, 2012). Le corpus multilingue OSCAR, basé sur du contenu moissonné à travers CommonCrawl [4], comprend également un sous-corpus occitan (Ortiz Suárez *et al.*, 2019). Enfin, un corpus basé sur des discussions Wikipédia a également été publié (Miletić & Scherrer, 2022a). Ce travail a été accompagnée d'expériences en identification de l'occitan dans un contenu multilingue (Miletić & Scherrer, 2022b).

Quant aux modèles de langue pré-entraînés, des plongements lexicaux pour l'occitan dérivés en utilisant fastText existent (Bojanowski *et al.*, 2017; Costa-jussà *et al.*, 2022), et l'occitan est également représenté dans le modèle contextualisé multilingue mBERT (Devlin *et al.*, 2019).

Notons que les expériences en étiquetage en parties du discours et en analyse syntaxique automatique citées ci-dessus ont été réalisés à l'aide des methodes d'apprentissage statistiques. À notre connaissance, en dehors d'expériences non encore publiées dans le cadre d'une thèse en cours sur la variation de l'occitan à l'épreuve des outils de TAL (Poujade, en préparation), il n'existe pas à ce jour de travaux sur l'occitan qui évalueraient l'efficacité des méthodes neuronales pour le traitement automatique de cette langue. Par ailleurs, nous n'avons pas identifié de travaux sur la lemmatisation de l'occitan. Or, ce niveau de traitement est crucial pour les langues à variation importante : comme la lemmatisation consiste à fournir pour chaque forme fléchie la forme de base correspondante, elle permet de réduire la dispersion de données dans un corpus et facilite ainsi l'apprentissage automatique.

3 Modification de ressources existantes

3.1 Tokénisation

Un tokéniseur à base de règles, implémenté en Perl, a été développé dans le cadre du projet RESTAURE pour le gascon et le languedocien (Vergez-Couret, 2019). Il met en place un traitement

3. Cf. Article 75-1 de la Constitution de la Cinquième République française.
4. https://commoncrawl.org/

spécial basé sur des listes d'unités polylexicales : elles ne sont pas tokénisées en formes graphiques individuelles, mais sont plutôt traitées en token unique (cf. *tre que* 'dès que' → [*tre_que*]).

La nouvelle version du tokéniseur est réalisée en Python. Nous reprenons les règles de base telles que définies dans la documentation de l'outil, mais abandonnons le traitement spécial des unités polylexicales. En effet, nous visons à produire une tokénisation plus compatible avec les exigences du projet Universal Dependencies [5], auxquelles se conforme également le corpus arboré Tolosa Treebank (Miletic *et al.*, 2020b). Comme le projet Universal Dependencies ne permet pas le maintien de ce type de tokens, nous effectuons une tokénisation en formes graphiques simples.

L'un des points principaux à traiter concerne le statut de l'apostrophe. À la différence du français, où l'apostrophe appartient systématiquement au token de gauche (cf. *l'apostrophe* → [*l', apostrophe*]), en occitan il peut appartenir aussi bien au token de gauche qu'à celui de droite. Par exemple, l'article défini se comporte d'une manière comparable au français : *l'acadèmia* → [*l', acadèmia*]. Le traitement diffère pour certains clitiques en post-position : *Ne'm vòs pas ?* 'Tu ne me veux pas ?' doit être tokénisée en [*Ne, 'm, vòs, pas, ?*]. Les différents cas de figure sont traités par des règles de tokénisation correspondantes, définies lors de la création du tokéniseur originel.

3.2 Conversion du lexique Loflòc vers le format UD

Loflòc (Lexic obèrt flechit occitan - Lexique ouvert fléchi occitan) (Vergez-Couret, 2016; Bras *et al.*, 2020) est un lexique morphosyntaxique développé dans le cadre du projet ANR RESTAURE puis du projet européen POCTEFA LINGUATEC, en collaboration avec Lo Congrès Permanent de la Lenga Occitana [6]. La version utilisée ici contient 849 605 entrées correspondant à 58 373 lemmes ; le contenu est en languedocien. Les informations morphosyntaxiques sont encodées à l'aide du jeu d'étiquettes GRACE (Rajman *et al.*, 1997). Il s'agit d'étiquettes positionnelles, où la première position encode la catégorie grammaticale, la deuxième la sous-catégorie sémantique, et les positions restantes les traits morphosyntaxiques pertinents pour la catégorie en question. Ainsi, l'étiquette `Ncfs` représente un nom (`N`) commun (`c`) féminin (`f`) singulier (`s`). La définition du jeu d'étiquettes pour l'occitan est disponible dans le manuel d'annotation du corpus étiqueté développé dans le cadre du projet RESTAURE [7].

Couplée à un corpus d'entraînement, cette ressource a le potentiel de faciliter la création de modèles performants pour le traitement automatique de l'occitan. Or, le corpus arboré Tolosa Treebank suit le schéma d'annotation du projet Universal Dependencies. Pour éliminer cette incompatibilité, nous avons effectué une conversion du lexique Loflòc vers le format UD. Le format UD encode les informations morphosyntaxiques à deux niveaux : les parties du discours sont exprimées en utilisant un jeu d'étiquettes réduit de 17 étiquettes [8], alors que les traits morphosyntaxiques sont encodés à l'aide de paires `Trait=Valeur` [9]. Notre script de conversion répartit les informations encodées dans les étiquettes GRACE entre ces deux niveaux. Une conversion GRACE → UD a été appliquée au premier corpus occitan étiqueté en parties du discours (Miletic *et al.*, 2019a), mais elle se limitait au niveau des étiquettes POS. Des exemples de conversion sont donnés dans le tableau 1 et la distribution d'entrées par partie du discours après la conversion est présentée dans le tableau 2.

5. https://universaldependencies.org/docs/u/overview/tokenization.html
6. https://locongres.org/fr/
7. https://doi.org/10.5281/zenodo.1173113
8. https://universaldependencies.org/u/pos/all.html
9. https://universaldependencies.org/u/feat/index.html

Forme	Lemme	GRACE	UD UPOS	UD Traits morphosyntaxiques
abausada	abausat	Afpfs	ADJ	Gender=Fem|Number=Sing|Degree=Pos
abscèsses	abscès	Ncmp	NOUN	Gender=Masc|Number=Plur
agirai	agir	Vmi-f1s-	VERB	Mood=Ind|Tense=Fut|Person=1|Number=Sing

TABLE 1 – Exemple de transformation GRACE → UD

Etiquette	Catégorie	Occurr.	Etiquette	Catégorie	Occurr.
ADJ	adjectif	58 995	NUM	numéral	231
ADP	adposition	670	PRON	pronom	519
ADV	adverbe	1 841	PROPN	nom propre	1 839
DET	déterminant	363	VERB	verbe	690 025
NOUN	nom commun	94 073			

TABLE 2 – Distribution d'entrées par partie du discours dans le lexique Loflòc au format UD

3.3 Ré-échantillonnage et correction de Tolosa Treebank

Le corpus arboré Tolosa Treebank contient 26 mille tokens annotés en parties du discours, en lemmes et en dépendances syntaxiques suivant le schéma d'annotation Universal Dependencies. Le languedocien était le premier dialecte à être annoté et il reste majoritaire dans le corpus, mais des textes en gascon, limousin et provençal ont également été intégrés. Tous les textes sont en graphie classique. La décision d'initalement limiter le corpus à un dialecte et à une graphie a été motivée par le besoin de pouvoir contrôler les différents niveaux de variation (Miletic *et al.*, 2020a). Le choix du languedocien est dû à sa position centrale dans le continuum dialectal : on peut s'attendre à ce que le contenu annoté en languedocien soit le plus utile pour le traitement des autres dialectes. La stratification par dialecte dans la version actuelle du corpus est donnée dans le tableau 3.

	Phrases	Tokens	Types	L. phrase	T = L
Tous dialectes	1 522	26 122	6 196	17,16	64,7
Gascon	255	4 170	1 429	16,35	64,5
Languedocien	1 113	19 315	4 499	17,35	65,1
Limousin	77	1 344	596	17,45	63,8
Provençal	77	1 293	583	16,79	61,4

TABLE 3 – Contenu de Tolosa Treebank par dialecte. L. phrase : longueur moyenne de la phrase en tokens. T = L : % de tokens identiques à leur lemme.

La première version de ce corpus à quatre dialectes a été utilisée pour des expériences en analyse syntaxique automatique trans-dialectale (Miletic *et al.*, 2020b). Dans ce but, le corpus avait été divisé en échantillons d'entraînement et de test, mais un échantillon de développement (*dev set*) n'a pas été créé car les méthodes utilisées dans ces expériences ne l'exigeaient pas. Comme il est en général requis par les méthodes neuronales pour l'évaluation des modèles intermédiaires durant l'entraînement, nous opérons un nouveau découpage du corpus en échantillons *train*, *test* et *dev* pour le languedocien et le gascon. Pour le limousin et le provençal, la quantité de données actuellement annotées ne permet pas la création d'un échantillon supplémentaire. Ce redécoupage du corpus

	train			test					dev				
	Phrases	Tokens	Types	Phrases	Tokens	Types	Inc.	Amb.	Phrases	Tokens	Types	Inc.	Amb.
Tous dialectes	1 196	20 551	5 292	202	3 179	1 054	22,11	28,18	124	2 392	1 009	16,39	31,77
Gascon	195	3 258	1 173	35	421	230	26,37	23,28	25	491	267	19,35	33,60
Languedocien	884	15 494	3 937	130	1 920	577	19,64	27,50	99	1 901	814	15,62	31,30
Limousin	56	919	434	16	413	211	27,76	31,76	-	-	-	-	-
Provençal	61	880	424	16	413	211	23,49	32,69	-	-	-	-	-

TABLE 4 – Division de Tolosa Treebank en échantillons d'entraînement, de développement et de test. Inc : % de tokens inconnus. Amb : % de tokens ambigus.

signifie que les résultats obtenus sur les versions différentes ne seront pas directement comparables.

Des informations quantitatives sur les différents échantillons, pour l'ensemble du corpus et par dialecte, sont proposées dans le tableau 4. Pour les échantillons *dev* et *test*, nous ajoutons le pourcentage de tokens inconnus et ambigus par rapport à l'échantillon *train* correspondant. Un token est considéré comme inconnu s'il est absent de l'échantillon *train*, alors qu'un token ambigu y figure, mais se trouve associé à plusieurs annotations possibles. Comme nos expériences se focalisent sur la lemmatisation, le pourcentage de tokens ambigus correspond à la part de tokens qui sont associés à plusieurs lemmes.

4 Expériences en lemmatisation

L'objectif des expériences présentées ici est d'identifier des stratégies utiles pour la lemmatisation de l'occitan. Les stratégies explorées concernent trois dimensions principales : le paradigme d'apprentissage (adaptation de l'entraînement séquentiel *vs* l'entraînement joint appliqués à l'étiquetage et à la lemmatisation), la taille et la nature de données d'entraînement (utilité d'un corpus limité annoté manuellement *vs* un corpus large annoté automatiquement) et la gestion de la variation dialectale (efficacité de modèles spécifiques à chaque dialecte *vs* un modèle global). Nous entraînons également des modèles d'étiquetage en POS afin d'évaluer les outils de lemmatisation sur des étiquettes POS fournies automatiquement. Les expériences discutées ici ont été mises en parallèle avec des expériences sur la lemmatisation du bas saxon (Miletić & Siewert, 2023), permettant d'identifier des difficultés et des stratégies communes, particulièrement pertinentes dans le contexte des langues non standardisées et peu dotées.

4.1 Outils utilisés

MaChAmp (van der Goot *et al.*, 2021) : cet outil permet d'effectuer l'apprentissage multi-tâche et le paramétrage (*fine-tuning*) d'un éventail de tâches du TAL, y compris l'étiquetage en parties du discours, la lemmatisation, l'analyse syntaxique automatique, la modélisation du langage masquée et la génération de texte. MaChAmp utilise un modèle contextualisé pré-entraîné comme encodeur et effectue le *fine-tuning* en fonction des tâches en aval qui lui sont demandées. Chaque tâche dispose d'un décodeur dédié qui permet d'effectuer les prédictions pour la tâche en question. L'outil permet également de réaliser un entraînement initial pour une tâche donnée, puis de ré-entraîner le modèle pour la même tâche sur un deuxième jeu de données. Nous avons exploité cette fonctionnalité dans nos expériences de lemmatisation. Par défaut, MaChAmp utilise les plongements lexicaux du modèle

pré-entraîné mBERT (Devlin *et al.*, 2019).

Stanza NLP (Qi *et al.*, 2020) : cette chaîne de traitement intègre actuellement de modèles de traitement pour 66 langues différentes (mais pas pour l'occitan). Elle comprend des modules de tokénisation, d'analyse des tokens multi-mots, de lemmatisation, d'étiquetage en parties du discours et en traits morphosyntaxiques, d'analyse syntaxique en dépendances et de reconnaissance des entités nommées. Comme Stanza permet l'entraînement de nouveaux modèles, nous avons utilisé son étiqueteur, basé sur un modèle biLSTM, et son lemmatiseur, qui intègre un modèle neuronal de séquence à séquence (*seq2seq*). L'étiqueteur permet d'utiliser des plongements lexicaux statiques ; nous avons tiré profit de cette fonctionnalité en utilisant ceux de fastText [10].

4.2 Préannotation d'un corpus plus large

Le corpus Tolosa Treebank est d'une taille relativement restreinte (26K tokens). Nous avons donc souhaité évaluer l'utilité d'un corpus plus large, qui serait annoté de manière automatique. Pour ce faire, nous nous sommes servie du corpus extrait du Wikipédia en occitan diffusé dans *Leipzig Corpora Collection* (Goldhahn *et al.*, 2012). La version téléchargeable de ce corpus contient 100K phrases correspondant à 2M de tokens [11]. Le contenu est segmenté en phrases, mais pas annoté.

Afin d'effectuer l'étiquetage en parties du discours et la lemmatisation de ce corpus, nous avons utilisé l'outil MaChAmp. Dans ces expériences initiales avec l'outil, nous avons opté pour l'entraînement indépendant des modèles d'étiquetage et de lemmatisation. Les modèles ont été appris sur l'échantillon d'entraînement du corpus Tolosa Treebank. Nous avons utilisé les plongements lexicaux par défaut.

L'étiqueteur en parties du discours a atteint l'exactitude de 92,26 % sur l'échantillon de test contenant les quatre dialectes. Le résultat le plus élevé a été obtenu sur le languedocien (92,97 %) et le plus bas sur le provençal (89,10 %). L'exactitude globale du lemmatiseur a atteint 89,30 %, le résultat le plus élevé étant 93,33 % sur le languedocien et le limousin, et le plus bas 88,6 % sur le gascon.

Pour annoter le corpus Wikipédia, nous l'avons d'abord tokénisé à l'aide du tokéniseur présenté dans la section 3.1. Si un token disposait d'une entrée non ambiguë dans le lexique Loflòc, nous avons retenu l'information disponible dans le lexique. Dans le cas contraire, nous avons fait appel aux modèles MaChAmp. Environ un tiers des tokens ont été annotés à partir du lexique.

4.3 Résultats et discussion

Cette section est dédiée à la discussion des résultats de nos expériences en lemmatisation. Nous indiquons systématiquement l'exactitude moyenne et la déviation standard obtenues à partir de trois tours d'entraînement avec des *random seeds* différents. Nous évaluons les outils sur l'ensemble des tokens, mais aussi sur les tokens inconnus et ambigus. Les résultats discutés ici ont été obtenus sur les échantillons *test*. Les résultats sur les échantillons *dev* sont disponibles dans l'annexe A.

Nous avons d'abord évalué les performances de plusieurs modèles globaux, entraînés et évalués sur le contenu en quatre dialectes. Les résultats sont présentés dans le tableau 5. La colonne *train* indique le corpus d'entraînement : TTB correspond à Tolosa Treebank, WIKI au corpus Wikipédia et COMB à la combinaison des deux. Dans le cas de MaChAmp, WIKI+TTB indique un premier entraînement sur

10. https://fasttext.cc/docs/en/crawl-vectors.html
11. https://corpora.uni-leipzig.de/en?corpusId=oci_wikipedia_2021

le corpus Wikipédia suivi d'un ré-entraînement sur le Tolosa Treebank. La colonne *tâche* précise si le modèle a été entraîné sur la lemmatisation seule ou sur l'étiquetage aussi. *Cond. entraîn.* et *Cond. test* indiquent respectivement les informations utilisées pour l'entraînement et pour l'évaluation. Les modèles qui s'appuient sur les étiquettes POS ont été évalués en utilisant l'annotation automatique produite par le modèle d'étiquetage entraîné sur le même corpus.

Dans ce scénario, les meilleurs résultats ont été obtenus par le modèle Stanza entraîné sur le Tolosa Treebank en s'appuyant sur les étiquettes POS (exactitude globale : 93,21 %). Quant à MaChAmp, le modèle le plus performant est celui entraîné sur le corpus Wikipédia et ré-entrâiné sur le Tolosa Treebank dans le paradigme d'apprentissage joint (exactitude globale : 92,16 %). Ces modèles dépassent ceux entraînés sans accès à l'étiquetage en POS, ce qui confirme encore une fois l'utilité des étiquettes POS pour la lemmatisation.

Globalement, les résultats de MaChAmp varient relativement peu : la différence entre les modèles le moins et le plus performant est de <1 %. Le passage du corpus plus petit au corpus plus large annoté automatiquement, ainsi que le ré-entraînement de ce deuxième modèle sur le corpus *gold* apportent tous les deux des améliorations, mais celles-ci restent limitées (<0,5 %). Quant à Stanza, les différences entre les modèles sont plus prononcées, les scores allant de 90,35 % pour le modèle entraîné sur le Tolosa Treebank sans utiliser les étiquettes POS à 93,21 % pour le modèle entraîné sur le même corpus en exploitant l'annotation morphosyntaxique. Il est intéressant que l'ajout du corpus Wikipédia au corpus Tolosa Treebank ici n'apporte pas d'amélioration (92,49 % vs 93,21 %). Il semblerait donc que pour MaChAmp la quantité de données d'entraînement est plus importante que la qualité de l'annotation, alors que pour Stanza ce serait l'inverse.

La question de la fiabilité de l'annotation semble particulièrement importante pour le traitement des tokens inconnus et ambigus : pour les deux outils, les meilleurs scores sur ces deux catégories ont été obtenus par les modèles entraînés sur le corpus *gold*. L'utilisation du corpus annoté automatiquement entraîne des pertes d'environ 4-5 % sur les tokens inconnus et d'environ 3-4 % sur les tokens ambigus pour MaChAmp, alors que pour Stanza les pertes sont respectivement d'environ 10 % et 4 %.

Dans un deuxième temps, nous avons évalué l'adaptation de différents modèles à chacun des dialectes (cf. tableau 9). Nous avons également développé des modèles MaChAmp ciblés en effectuant le ré-entraînement sur chacun des sous-corpus (cf. les modèles WIKI+GA, WIKI+LI, WIKI+LA, WIKI+PR) pour identifier la meilleure stratégie pour les dialectes individuels.

Le modèle Stanza entraîné sur le Tolosa Treebank reste le plus utile : il atteint les meilleurs résultats sur les trois catégories de tokens pour le gascon, le languedocien et le provençal, et il est également le plus performant sur les tokens ambigus sur l'échantillon limousin. Pour ce dernier dialecte, c'est le modèle MaChAmp ré-entraîné sur l'ensemble du corpus Tolosa Treebank qui gagne sur les tokens inconnus et sur l'ensemble des tokens. Pour les quatre dialectes, le ré-entraînement sur les sous-corpus dédiés donne des résultats moins élevés que l'utilisation de l'ensemble du corpus Tolosa Treebank. Les pertes sont les plus prononcées sur le limousin et le provençal (respectivement d'environ 3 % et 5 %). Comme ces deux dialectes disposent des sous-corpus les plus petits, il est possible que cet effet soit dû à la taille de l'échantillon de ré-entraînement.

Globalement, l'utilité d'un corpus large annoté automatiquement semble dépendre de l'outil : avec MaChAmp, le corpus plus large apporte une amélioration au niveau du modèle global, mais ce n'est pas le cas de Stanza. Cet outil semble favoriser la fiabilité de l'annotation et obtient les meilleurs résultats à partir du corpus *gold*. Plus généralement, sur les token inconnus, les modèles entrâinés sur le corpus *gold* surpassent les modèles entraînés sur le corpus plus large. Si l'on vise à optimiser

Outil	*train*	Tâche	Cond. entraîn.	Cond. test	Tous	Inconnus	Ambigus
MaChAmp	TTB	LEM	no POS, gold LEM	no POS	$91,28^{\pm 0,42}$	$72,22^{\pm 1,55}$	$96,23^{\pm 0,37}$
	WIKI	POS+LEM	pred. POS+LEM	no POS	$91,77^{\pm 0,23}$	$68,54^{\pm 1,86}$	$92,19^{\pm 0,14}$
	WIKI+TTB	POS+LEM	pred. POS+LEM	no POS	$92,16^{\pm 0,25}$	$67,20^{\pm 0,33}$	$93,05^{\pm 0,45}$
Stanza	TTB	LEM	no POS, gold LEM	no POS	$90,35^{\pm 0,42}$	$66,86^{\pm 1,85}$	$95,78^{\pm 0,00}$
	TTB	LEM	gold POS+LEM	pred. POS	**93,21** $^{\pm 0,09}$	**78,43** $^{\pm 0,41}$	**96,69** $^{\pm 0,00}$
	COMB	LEM	pred. POS+LEM	pred. POS	$92,49^{\pm 0,08}$	$68,40^{\pm 0,98}$	$92,63^{\pm 0,00}$

TABLE 5 – Lemmatisation : résultats sur l'ensemble de l'échantillon de test

		Gascon				**Limousin**		
Outil	*train*	Tous	Inconnus	Ambigus	*train*	Tous	Inconnus	Ambigus
MaChAmp	WIKI+TTB	$89,66^{\pm 0,52}$	$57,01^{\pm 1,24}$	$90,28^{\pm 0,57}$	WIKI+TTB	**90,91** $^{\pm 0,20}$	**74,42** $^{\pm 1,90}$	$94,35^{\pm 0,46}$
	WIKI+GA	$88,86^{\pm 0,41}$	$54,38^{\pm 1,24}$	$89,58^{\pm 0,98}$	WIKI+LI	$87,64^{\pm 0,57}$	$64,34^{\pm 1,10}$	$92,66^{\pm 0,80}$
Stanza	TTB	**90,71** $^{\pm 0,75}$	**77,78** $^{\pm 2,79}$	**91,49** $^{\pm 0,00}$	TTB	$90,59^{\pm 0,41}$	$72,60^{\pm 0,80}$	**99,22** $^{\pm 0,00}$
	COMB	$90,06^{\pm 0,11}$	$67,54^{\pm 1,24}$	$89,58^{\pm 0,00}$	COMB	$89,79^{\pm 0,23}$	$66,67^{\pm 1,09}$	$92,66^{\pm 0,00}$
		Languedocien				**Provençal**		
Outil	*train*	Tous	Inconnus	Ambigus	*train*	Tous	Inconnus	Ambigus
MaChAmp	WIKI+TTB	$93,08^{\pm 0,48}$	$69,91^{\pm 0,33}$	$92,76^{\pm 0,69}$	WIKI+TTB	$91,67^{\pm 0,00}$	$54,67^{\pm 1,89}$	$95,14^{\pm 0,44}$
	WIKI+LA	$92,56^{\pm 0,60}$	$68,29^{\pm 0,33}$	$92,29^{\pm 0,80}$	WIKI+PR	$86,60^{\pm 0,11}$	$52,00^{\pm 0,00}$	$89,55^{\pm 0,25}$
Stanza	TTB	**94,42** $^{\pm 0,13}$	**81,35** $^{\pm 0,90}$	**96,54** $^{\pm 0,00}$	TTB	**92,81** $^{\pm 0,31}$	**74,92** $^{\pm 1,28}$	**98,51** $^{\pm 0,00}$
	COMB	$93,72^{\pm 0,11}$	$71,53^{\pm 1,50}$	$92,98^{\pm 0,00}$	COMB	$92,08^{\pm 0,12}$	$54,67^{\pm 1,89}$	$93,51^{\pm 0,00}$

TABLE 6 – Lemmatisation : résultats par dialecte

les performances des modèles sur cette catégorie de tokens, il semble utile de favoriser la qualité de l'annotation plutôt que la taille du corpus d'entraînement. Enfin, avec MaChAmp, les entraînements ciblés par dialecte n'apportent pas d'amélioration par rapport à l'entraînement global. Néanmoins, la taille de nos sous-corpus étant encore limitée, notamment pour le limousin et le provençal, cette stratégie mériterait d'être ré-évaluée sur des versions futures du corpus.

5 Corpus Wikipédia annoté

Afin de favoriser les travaux sur l'occitan, nous distribuons le corpus Wikipédia étiqueté en parties du discours et lemmatisé dans le cadre de nos expériences. Les 2 millions de tokens correspondent à 111 656 lemmes différents. La distribution des parties du discours se trouve dans le tableau 7. Dans la version actuelle du corpus, nous gardons l'annotation initiale (voir section 4.2). Celle-ci évoluera à l'avenir afin de prendre en compte les résultats présentés ici. Le corpus est disponible à l'adresse suivante : https://doi.org/10.5281/zenodo.7777340.

6 Conclusion

Nous avons présenté plusieurs avancées dans les efforts pour outiller l'occitan, une langue régionale non standardisée. Une part de ces efforts était focalisée sur des ressources existantes : des améliorations mineures ont été apportées à un tokéniseur existant et au corpus arboré Tolosa Treebank, alors

Etiquette	Catégorie	Occ.	Etiquette	Catégorie	Occ.
ADJ	adjectif	135 760	NUM	numéral	45 116
ADP	adposition	306 798	PART	particule	4 610
ADV	adverbe	76 063	PRON	pronom	65 916
AUX	verbe auxiliaire	55 023	PROPN	nom propre	92 554
CCONJ	conj. de coord.	56 173	PUNCT	ponctuation	225 596
DET	déterminant	317 677	SCONJ	subordonnant	20 831
INTJ	interjection	3 589	VERB	verbe	191 342
NOUN	nom commun	434 306	X	foreign	1 654

TABLE 7 – Distribution de parties du discours dans le corpus Wikipédia

que le lexique Loflòc a été converti du format GRACE vers le format Universal Dependencies. Deuxièmement, nous avons présenté, à notre connaissance, la première évaluation en lemmatisation de l'occitan. Ces expériences ont été réalisées à l'aide d'outils neuronaux, qui ne figurent pas pour l'heure dans les travaux publiés sur le traitement de l'occitan. Nous avons également décrit une nouvelle ressource : un corpus extrait du Wikipédia, qui a été tokénisé, étiqueté en parties du discours et lemmatisé. Ce corpus est désormais disponible sur Zenodo.

Les modèles de lemmatisation que nous avons entraînés atteignent des résultats solides, avec un taux d'exactitude allant de 90,71 % sur le gascon jusqu'à 94,42 % sur le languedocien, l'exactitude globale sur l'ensemble des dialectes étant de 93,21 %. Comme nous avons entraîné tous les modèles avec les valeurs des paramètres par défaut, il est possible que des tests de paramétrage apportent des gains supplémentaires. L'utilisation d'un corpus large annoté automatiquement a amélioré les résultats avec l'outil MaChAmp ; en revanche, Stanza s'est montré plus performant en limitant l'entraînement aux données annotées manuellement. Quant aux entraînements ciblés pour chaque dialecte, nous n'avons pas observé d'effets positifs ; qui plus est, cette stratégie entraîne des pertes importantes pour le limousin et le provençal. Ces observations sont sans doute liées à la taille limitée des sous-corpus individuels ; elles méritent d'être ré-évaluées sur les futures versions du corpus Tolosa Treebank.

Nous espérons que ces résultats, ainsi que les ressources présentées ici, favoriseront la poursuite des travaux sur l'occitan et faciliteront sa sortie du statut de langue peu dotée.

Remerciements

Je souhaite remercier Myriam Bras (Université de Toulouse) et Marianne Vergez-Couret (Université de Poitiers) pour leurs retours précieux.

Ce travail a été effectué dans le cadre du projet "CorCoDial – Corpus-based computational dialectology" (Academy of Finland, No. 342859).

A Résultats sur l'échantillon *dev*

Outil	*train*	Tâche	Cond. entraîn.	Cond. test	Tous	Inconnus	Ambigus
MaChAmp	TTB	LEM	no POS, gold LEM	no POS	$93,57^{\pm 0,06}$	$78,74^{\pm 1,14}$	$95,08^{\pm 0,28}$
	WIKI	POS+LEM	pred. POS+LEM	no POS	$93,32^{\pm 0,09}$	$76,07^{\pm 0,50}$	$91,94^{\pm 0,15}$
	WIKI+TTB	POS+LEM	pred. POS+LEM	no POS	$94,24^{\pm 0,17}$	$73,49^{\pm 0,74}$	$93,47^{\pm 0,29}$
Stanza	TTB	LEM	no POS, gold LEM	no POS	$92,84^{\pm 0,14}$	$75,43^{\pm 0,84}$	$93,10^{\pm 0,0}$
	TTB	LEM	gold POS+LEM	pred. POS	$94,68^{\pm 0,03}$	$83,16^{\pm 0,21}$	$94,86^{\pm 0,0}$
	COMB	LEM	pred. POS+LEM	pred. POS	$93,53^{\pm 0,06}$	$74,01^{\pm 1,11}$	$91,19^{\pm 0,0}$

TABLE 8 – Exactitude de lemmatisation sur tous les dialectes. Échantillon *dev*.

		Gascon				**Languedocien**		
Outil	*train*	Tous	Inconnus	Ambigus	*train*	Tous	Inconnus	Ambigus
MaChAmp	WIKI+TTB	$93,60^{\pm 0,36}$	$69,10^{\pm 1,15}$	$93,55^{\pm 1,11}$	WIKI+S	$94,40^{\pm 0,14}$	$75,58^{\pm 0,95}$	$93,45^{\pm 0,1}$
	WIKI+GA	$92,83^{\pm 0,10}$	$69,10^{\pm 2,30}$	$92,14^{\pm 0,22}$	WIKI+LA	$94,12^{\pm 0,13}$	$74,03^{\pm 1,09}$	$93,25^{\pm 0,11}$
Stanza	WIKI	$94,29^{\pm 0,20}$	$79,65^{\pm 0,99}$	$96,15^{\pm 0,00}$	TTB	$94,78^{\pm 0,02}$	$84,29^{\pm 0,16}$	$94,51^{\pm 0,0}$
	COMB	$90,40^{\pm 0,00}$	$68,29^{\pm 0,00}$	$87,26^{\pm 0,00}$	COMB	$94,33^{\pm 0,08}$	$76,75^{\pm 1,65}$	$92,16^{\pm 0,0}$

TABLE 9 – Exactitude de lemmatisation par dialecte. Échantillon *dev* (le corpus utilisé propose des échantillons *dev* seulement pour le gascon et le languedocien.)

Références

BEC P. (1995). *La langue occitane*. PUF, 6th édition.

BERGMANIS T. & GOLDWATER S. (2018). Context sensitive neural lemmatization with Lematus. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)*, p. 1391–1400, New Orleans, Louisiana : Association for Computational Linguistics. DOI : 10.18653/v1/N18-1126.

BERNHARD D., LIGOZAT A.-L., BRAS M., MARTIN F., VERGEZ-COURET M., ERHART P., SIBILLE J., TODIRASCU A., BOULA DE MAREÜIL P. & HUCK D. (2021). Collecting and annotating corpora for three under-resourced languages of france : Methodological issues. *Language Documentation & Conservation*, (15), 316–357.

BERNHARD D., LIGOZAT A.-L., MARTIN F., BRAS M., MAGISTRY P., VERGEZ-COURET M., STEIBLÉ L., ERHART P., HATHOUT N., HUCK D., REY C., REYNÉS P., ROSSET S., SIBILLE J. & LAVERGNE T. (2018). Corpora with part-of-speech annotations for three regional languages of France : Alsatian, Occitan and Picard. In *International Conference on Language Resources and Evaluation*, Miyazaki, Japan. HAL : hal-02358018.

BOJANOWSKI P., GRAVE E., JOULIN A. & MIKOLOV T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, **5**, 135–146. DOI : 10.1162/tacl_a_00051.

BRAS M. & VERGEZ-COURET M. (2016). BaTelÒc : a text base for the Occitan language. In V. FERREIRA & P. BOUDA, Éds., *Language Documentation and Conservation in Europe*, p. 133–149 : Honolulu : University of Hawaï Press. HAL : hal-00987241.

BRAS M., VERGEZ-COURET M., HATHOUT N., SIBILLE J., SÉGUIER A. & DAZÉAS B. (2020). Lofl̀oc : Lexic obèrt flechit occitan. In J.-F. COUROUAU, Éd., *Fidélités et dissidences (Actes du XIIe congrès de l'Association Internationale d'Études Occitanes)*, Albi : Centre d'Etude de la Littérature Occitane.

CORRAL A., LETURIA I., SÉGUIER A., BARRET M., DAZÉAS B., BOULA DE MAREÜIL P. & QUINT N. (2020). Neural text-to-speech synthesis for an under-resourced language in a diglossic environment : the case of Gascon Occitan. In *Proceedings of the 1st Joint SLTU (Spoken Language Technologies for Under-resourced languages) and CCURL (Collaboration and Computing for Under-Resourced Languages) Workshop «Language Resources and Evaluation Conference–Marseille–11–16 May 2020»*, p. 53–60 : European Language Resources Association (ELRA).

COSTA-JUSSÀ M. R., CROSS J., ÇELEBI O., ELBAYAD M., HEAFIELD K., HEFFERNAN K., KALBASSI E., LAM J., LICHT D., MAILLARD J. *et al.* (2022). No language left behind : Scaling human-centered machine translation. *arXiv preprint arXiv :2207.04672*.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

GESMUNDO A. & SAMARDZIC T. (2012). Lemmatising serbian as category tagging with bidirectional sequence classification. In N. C. C. CHAIR), K. CHOUKRI, T. DECLERCK, M. U. DOĞAN, B. MAEGAARD, J. MARIANI, A. MORENO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey : European Language Resources Association (ELRA).

GOLDHAHN D., ECKART T. & QUASTHOFF U. (2012). Building large monolingual dictionaries at the Leipzig corpora collection : From 100 to 200 languages. In N. CALZOLARI, K. CHOUKRI, T. DECLERCK, M. U. DOĞAN, B. MAEGAARD, J. MARIANI, A. MORENO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey : European Language Resources Association (ELRA).

GRZEGORZ CHRUPALA G. D. & VAN GENABITH J. (2008). Learning morphology with morfette. In B. M. J. M. J. O. S. P. D. T. NICOLETTA CALZOLARI (CONFERENCE CHAIR), KHALID CHOUKRI, Éd., *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco : European Language Resources Association (ELRA). http ://www.lrec-conf.org/proceedings/lrec2008/.

MANJAVACAS E., KÁDÁR Á. & KESTEMONT M. (2019). Improving lemmatization of non-standard languages with joint learning. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 1493–1503, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1153.

MILETIC A., BERNHARD D., BRAS M., LIGOZAT A.-L. & VERGEZ-COURET M. (2019a). Transformation d'annotations en parties du discours et lemmes vers le format universal dependencies : étude de cas pour l'alsacien et l'occitan. Poster, HAL : hal-02123743.

MILETIC A., BRAS M., ESHER L., SIBILLE J. & VERGEZ-COURET M. (2019b). Building a treebank for Occitan : what use for Romance UD corpora ? In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, p. 2–11, Paris, France : Association for Computational Linguistics. DOI : 10.18653/v1/W19-8002.

MILETIC A., BRAS M., VERGEZ-COURET M., ESHER L., POUJADE C. & SIBILLE J. (2020a). Building a Universal Dependencies Treebank for Occitan. In *Proceedings of The 12th Language Resources and Evaluation Conference*, p. 2932–2939, Marseille, France : European Language Resources Association.

MILETIC A., BRAS M., VERGEZ-COURET M., ESHER L., POUJADE C. & SIBILLE J. (2020b). A four-dialect treebank for Occitan : Building process and parsing experiments. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, p. 140–149, Barcelona, Spain (Online) : International Committee on Computational Linguistics (ICCL).

MILETIĆ A. & SIEWERT J. (2023). Lemmatization experiments on two low-resourced languages : Low Saxon and Occitan. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, p. 163–173, Dubrovnik, Croatia : Association for Computational Linguistics.

MILETIĆ A. & SCHERRER Y. (2022a). OcWikiDisc : a Corpus of Wikipedia Talk Pages in Occitan. DOI : 10.5281/zenodo.7079580.

MILETIĆ A. & SCHERRER Y. (2022b). OcWikiDisc : a corpus of Wikipedia talk pages in Occitan. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, p. 70–79, Gyeongju, Republic of Korea : Association for Computational Linguistics.

MÜLLER T., COTTERELL R., FRASER A. & SCHÜTZE H. (2015). Joint lemmatization and morphological tagging with lemming. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, p. 2268–2274, Lisbon, Portugal : Association for Computational Linguistics. DOI : 10.18653/v1/D15-1272.

ORTIZ SUÁREZ P. J., SAGOT B. & ROMARY L. (2019). Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. In P. BAŃSKI, A. BARBARESI, H. BIBER, E. BREITENEDER, S. CLEMATIDE, M. KUPIETZ, H. LÜNGEN & C. ILIADI, Éds., *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019*, p. 9 – 16, Mannheim : Leibniz-Institut für Deutsche Sprache. DOI : 10.14618/ids-pub-9021.

POUJADE C. (en préparation). *La linguistique outillée à l'épreuve de la variation : Ressources et outils pour les parlers occitans de l'Ariège*. Thèse de doctorat, Université de Toulouse.

QI P., ZHANG Y., ZHANG Y., BOLTON J. & MANNING C. D. (2020). Stanza : A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics : System Demonstrations*.

RAJMAN M., LECOMTE J. & PAROUBEK P. (1997). *Format de description lexicale pour le français. Partie 2 : Description morpho-syntaxique*. Rapport interne, EPFL & INaLF. GRACE GTR-3-2.1.

SIBILLE J. (2002). Ecrire l'occitan : essai de présentation et de synthèse. In D. CAUBET, S. CHAKER & J. SIBILLE, Éds., *Les langues de France et leur codification. Ecrits divers – Ecrits ouverts*, Paris, France : Inalco / Association Universitaire des Langues de France L'Harmattan. HAL : hal-01296986.

VAN DER GOOT R., ÜSTÜN A., RAMPONI A., SHARAF I. & PLANK B. (2021). Massive choice, ample tasks (MaChAmp) : A toolkit for multi-task learning in NLP. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics : System Demonstrations*, p. 176–197, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.eacl-demos.22.

VERGEZ-COURET M. (2016). *Description du lexique Lofl̀oc*. Research report, CLLE-ERSS. HAL : hal-01338774.

VERGEZ-COURET M. (2019). Tokenization for occitan (gascon and lengadocian). DOI : 10.5281/zenodo.2533873.

VERGEZ-COURET M. & URIELI A. (2015). Analyse morphosyntaxique de l'occitan languedocien : l'amitié entre un petit languedocien et un gros catalan. In *TALARE 2015*, Caen, France.