Analyse morphosyntaxique de l'occitan languedocien : l'amitié entre un petit languedocien et un gros catalan

Marianne Vergez-Couret[1] Assaf Urieli[1,2]

(1) CLLE-ERSS, CNRS, Université de Toulouse 2, 5 allées Antonio Machado, 31058 TOULOUSE cedex 9
(2) Joliciel Informatique SARL, 2 avenue du Cardié, 09000 FOIX
marianne.vergez@univ-tlse2.fr, assaf.urieli@univ-tlse2.fr

Résumé. Dans cette étude, nous nous intéressons à la question de l'analyse morphosyntaxique de l'occitan. Nous utilisons Talismane, un logiciel par apprentissage supervisé, nécessitant des données annotées pour l'entraînement et optionnellement un lexique. Nous montrons dans cet article**,** qu'en l'absence de données annotées suffisantes pour l'occitan, il est possible d'obtenir de bons résultats (92%) en utilisant les données d'une langue étymologiquement proche, le catalan. Nous avons utilisé le corpus Ancora (500 000 formes) et un lexique occitan languedocien (250 000 entrées). Utiliser un corpus catalan de taille importante permet une amélioration de +3% par rapport au résultat obtenu avec le seul corpus d'entraînement occitan disponible à ce jour de 2800 formes.

Abstract.

Pos-tagging the Lengadocian dialect of Occitan: a little Lengadocian befriends a big Catalan.
In this study, we examine the question of Occitan POS-tagging. We use Talismane, a supervised machine learning NLP tool, requiring annotated data for training and optionally a lexicon. We show that, with insufficient data for Occitan, it is possible to obtain good results (92%) by using data from an etymologically close language, in this case Catalan. We used the Catalan Ancora corpus (500,000 tokens) and an Occitan Languedocien lexicon (250,000 entries). Using the larger Catalan corpus improved results by +3% with respect to the result obtained using the only Occitan training corpus available to date (2,800 tokens).

Mots-clés : traitement automatique des langues peu dotées, occitan, analyse morphosyntaxique

Keywords: natural language processing for lesser resourced languages, Occitan, POS-tagging.

1 Introduction

Les méthodes les plus couramment employées pour développer des outils de TAL sont à l'heure actuelle des méthodes par apprentissage supervisé quand des données annotées sont disponibles. Nous inscrivons nos travaux dans cette tendance pour l'analyse morphosyntaxique automatique de l'occitan. Il est donc nécessaire de rassembler des lexiques et des corpus annotés. Construire ces ressources requiert des efforts conséquents et des moyens financiers et humains qui font souvent défaut dans le cas des langues peu dotées[1] comme l'occitan.
Dans cet article, nous proposons de comparer les résultats obtenus avec un corpus d'entraînement languedocien de petite taille (2800 formes) et un corpus d'entraînement catalan de grande taille (500 000 formes). Le corpus languedocien a l'avantage d'avoir été annoté par nos soins et correspond parfaitement aux besoins d'annotation attendus. Il favorise la qualité au détriment de la quantité. Le corpus catalan que nous utilisons pour cette étude est Ancora (Taulé et al., 2008), un corpus de 500 000 formes annotées. Nous souhaitons évaluer s'il faut, dans la constitution ou l'exploitation de ressources pour l'entraînement d'un analyseur morphosyntaxique, favoriser la qualité ou la quantité des annotations. Pour ce faire, nous avons mis en place plusieurs expériences : a) entraînement avec chacun des deux corpus séparés (après une étape d'harmonisation) ; b) entraînement avec transformation superficielle du corpus catalan avec transposition en occitan des mots les plus fréquents ; c) entraînement avec combinaison des deux corpus en faisant varier le poids attribué au corpus languedocien.

[1] Le terme "langues peu dotées" pour les langues disposant de peu ou pas de ressources et d'outils linguistiques informatisées sera utilisé en opposition à "langues très dotées" pour celles qui disposent d'un grand nombre de ressources et d'outils.

En section 1, nous présentons l'occitan et le catalan ainsi que les principaux traits de ressemblance et de différence lexicale, morphologique et syntaxique des deux langues. Puis, nous présentons en section 2 les principes de fonctionnement de Talismane (Urieli, 2013) que nous avons choisi pour cette étude avant d'aborder plus généralement en section 3 quelques méthodes courantes en traitement automatique des langues peu dotées. Nous présentons, en section 4, les ressources que nous avons rassemblées pour cette étude et en section 5 les expériences que nous avons menées ainsi que les résultats que nous avons obtenus.

2 Deux langues étymologiquement proches : Occitan et Catalan

L'occitan et le catalan font partie de l'ensemble des langues romanes. Elles sont toutes deux issues de la fragmentation dialectale de l'ensemble gallo-roman méridional. Pierre Bec (1995) divise cet ensemble en quatre *complexus dialectaux :* le nord-occitan (limousin, auvergnat, vivaro-alpin), l'occitan méridional (languedocien, provençal), le gascon et le catalan. Il faut souligner que l'appartenance du catalan à cet ensemble gallo-roman est une position idéologique qui n'est pas partagée par tous. Néanmoins, P. Bec explique cela par les «extraordinaires ressemblances" que l'occitan et le catalan présentent. Toutefois, les deux langues, bien que très proches à l'origine, furent séparées politiquement et culturellement à partir du 13ème siècle comme le souligne Sibille (1996). Bec (1995) distingue le catalan de l'occitan méridional, et en particulier du languedocien, par des traits phonétiques (par ailleurs répercutés sur la graphie) et une certaine originalité du lexique. A l'heure actuelle, l'occitan et le catalan ont des situations sociolinguistiques bien différentes. Nous résumons ci-dessous quelques caractéristiques de ces deux langues.

2.1 Occitan

L'occitan est parlé dans le sud de la France et dans quelques vallées espagnoles et catalanes. Il est difficile d'estimer le nombre de locuteurs occitans mais plusieurs études permettent d'établir un chiffre aux alentours de 500 000 locuteurs. Néanmoins, un nombre plus important de locuteurs, estimés aux alentours de 1,5 millions de personnes[2], peuvent manifester un intérêt pour la langue occitane (la pratiquer et/ou la comprendre avec divers degrés de compétences). Cet intérêt pour l'occitan est appuyé par un important réseau associatif parmi lequel les écoles primaires et secondaires en immersion bilingue Calandreta, l'Institut d'Estudis Occitan (IEO), le CFPO (Centre de Formacion Professionala Occitan) qui offrent des formations en occitan pour tous les âges. L'occitan est également présent dans le système éducatif français sous forme d'options obligatoires ou facultatives du primaire à l'université.

2.2 Catalan

Le catalan est parlé en Catalogne autonome, en Catalogne Nord (Pyrénées Orientales), en Andorre, en Pays Valencien, dans les îles Baléares, dans la Frange orientale d'Aragon, dans la ville d'Alguer (en Sardaigne) et dans la région du Carxe en Murcie, regroupant 9,6 millions de locuteurs sur un total de 13,5 millions d'habitants (Almarcha París et Baylac Ferrer, 2012). La pratique courante de la langue catalane dans l'enseignement, les médias, la politique et les entreprises place le catalan dans une position intermédiaire entre langues normalisées (comme le castillan) et langues minorisées (comme l'occitan), bien qu'Almarcha París et Baylac Ferrer (2012) signalent des disparités de cette pratique selon les régions concernées. La quantité de données linguistiques disponibles pour le catalan qui occupe une place de choix sur internet (Serra Serra, 2012) est également un indice fort de cette position (Boleda *et al*, 2009). Selon nous, le catalan est dans l'ensemble des langues romanes, la langue la plus proche de l'occitan. Nous souhaitons donc mettre en place une méthode permettant de tirer parti au mieux des ressources linguistiques disponibles pour le catalan (dans notre cas le corpus Ancora[3]) pour une tâche d'analyse morphosyntaxique de la variante languedocienne de l'occitan. Nous allons présenter dans la section suivante des traits similaires et différents des deux langues.

2.3 Ressemblances et différences lexicales, morphologiques, syntaxiques

Nous allons décrire dans les sections suivantes quelques ressemblances et différences majeures de la variante languedocienne de l'occitan et du catalan standard sans toutefois viser l'exhaustivité.

[2] D'après une étude socio-linguistique réalisée en Midi-Pyrénées en 2010 (http://www.midipyrenees.fr/IMG/pdf/EnqueteOccitan.pdf)

[3] De nombreux outils de TAL (FreeLing) et des lexiques (Apertium) sont également disponibles librement pour le catalan et pourront être exploités pour d'autres expériences.

2.3.1 Caractéristiques phonétiques et lexicales

Les principales distinctions phonétiques en languedocien et en catalan sont décrites dans (Bec, 1995). Elles sont le résultat d'évolutions divergentes qui ont conduit à des distinctions phonétiques et également graphiques.

Latin	Clave	Cantare	Capra	Causa	Lingua
Catalan	Clau	Cantar	Cabra	Cosa	Llengua
Occitan	Clau	Cantar	Cabra	Causa	Lenga

Tableau 1. Mots occitans et catalans d'origine latine

Par exemple, l'occitan a conservé la diphtongue *au* et pas le catalan *causa/cosa*. Le couple *llengua/lenga* illustre la palatalisation du l- initial en catalan. Bec signale également une certaine originalité du lexique du catalan, par exemple pour les mots proches du castillan comme *molt.*

2.3.2 Ressemblances et différences morphologiques

Tandis que le masculin pluriel et le féminin singulier réguliers des adjectifs catalans et occitans sont similaires, les tableaux ci-dessous montrent la différence pour le féminin pluriel.

	Masculin/Singulier	Masculin/Pluriel	Féminin/Singulier	Féminin/Pluriel
Catalan	Madur	Madurs	Madura	Madur**es**
Occitan	Madur	Madurs	Madura	Madur**as**

Tableau 2. Adjectif *Madur (mûr)* en catalan et en occitan

De nombreuses différences peuvent également être relevées dans les flexions verbales de l'occitan et du catalan.

Catalan	Cant**o**	Cant**es**	Canta	Cant**em**	Cant**eu**	Cant**en**
Occitan	Cant**i**	Cant**as**	Canta	Cant**am**	Cant**atz**	Cant**an**

Tableau 3. Verbe cantar (chanter) présent de l'indicatif en catalan et en occitan

2.3.3 Différences syntaxiques

Nous proposons dans cette section de dégager quelques caractéristiques de l'occitan et du catalan à partir des exemples suivants en languedocien (pour les parties a) et en catalan (pour les parties b) de la *Parabole de l'Enfant prodigue* repris de Bec (1995) en nous focalisant sur celles qui ont un impact sur l'ordre et la distribution des mots dans les deux langues.

La négation ne se marque pas de la même façon en catalan et en languedocien. En catalan, la marque de la négation se place avant le verbe, cf «només tenia" en 1 et en «no va tenir" en 2 tandis qu'en languedocien la ou les marques de négation seront exprimées après le verbe «aviá pas que" en 1 et «aguèt pas mei" en 2.

1. *a) Un òme aviá **pas que** dos dròlles.*
   *b) Un home **només** tenia dos fills.*
   'Un homme n'avait que deux fils.'

2. *a) **Aguèt pas mei** de lèit per dormir la nuèit ni de fuòc per **se calfar** quand aviá freg.*
   *b) Ja **no va tenir** llit per a dormir a la nit ni foc per a **escalfar-se** quan tenia fred.*
   'Il n'eut plus de lit pour dormir la nuit ni de feu pour se chauffer quand il avait froid.'

Il existe une différence majeure en catalan et en languedocien concernant la structure *anar* (aller) + verbe à l'infinitif qui en catalan permet d'exprimer le prétérit comme en 3 : «va dir" et en languedocien permet d'exprimer le futur proche.

3. *a) Lo plus jove **diguèt a son** paire :*
   *b) El més jove **va dir al seu** pare :*
   'Le plus jeune dit à son père :'

Mais au final, cette différence interprétative n'a aucune incidence sur l'annotation morphosyntaxique que nous proposons qui sera la même en catalan et en languedocien :

va anar Vc-Pri-P3-sg (Verbe conjugué au présent de l'indicatif, troisième personne du singulier)
cantar cantar Vi (Verbe à l'infinitif)

La position des pronoms en catalan et en languedocien est un problème complexe et sujet à beaucoup de variations. Toutefois, dans les phrases conjuguées, les pronoms se placent avant le verbe, par exemple en 4 « *li balhava* » en languedocien et « *li donava* » en catalan. L'ordre des pronoms peut changer en catalan et en languedocien : « *me la compro* » vs. « *la me crompi* » mais ce phénomène aura probablement peu de conséquences pour notre tâche d'annotation morphosyntaxique étant donné que les deux pronoms auront la même catégorie grammaticale (pronom clitique). En revanche, en catalan standard, le pronom apparaît toujours après le verbe lorsqu'il est à l'infinitif, par exemple « *escalfar-se* » (« *se calfar* » en languedocien) en 5 et « *anar-me'n* » en 6, ce qui est impossible en languedocien.

4. *a) Mas degun **li balhava** pas res.*
   *b) Pero ningú no **li donava** res.*
   'Mais personne ne lui donnait rien.'

5. *a) **Aguèt pas mei** de lèit per dormir la nuèit ni de fuòc per **se calfar** quand aviá freg.*
   *b) Ja **no va tenir** llit per a dormir a la nit ni foc per a **escalfar-se** quan tenia fred.*
   'Il n'eut plus de lit pour dormir la nuit ni de feu pour se chauffer quand il avait froid.'

6. *a) Es ora per ièu de me governar sol e d'aver argent ; **me** cal poder partir e véser de païs.*
   *b) Ja és hora que sigui el meu propi amo i que tingui diners ; cal que pugui **anar-me'n** i veure món.*
   'Il est temps pour moi d'être indépendant et de gagner de l'argent ; il faut que je parte et que je voie du pays.'

Il existe en occitan et en catalan plusieurs façons de marquer la possession. En catalan standard, on retrouve principalement les formes composées (*el meu, la meva, els meus, les meves, el nostre, la nostra, els nostres, les nostres*…) bien qu'il existe également des formes simples (*ma, ta, sa, vostre, nostre, son*). En 3, une forme simple en languedocien « *a son paire* » est traduite par une forme composée « *al (a+el) seu pare* » en catalan tandis qu'en 7 on trouve une forme composée dans les deux cas : « *lo vòstre ben* » et « *el vostre bé* » . Il est également possible de trouver un adjectif possessif après le nom : tandis que ce sera un phénomène rare en catalan illustré en 8, « *fill meu* », il permet avec plus de liberté de marquer une insistance en languedocien : « *l'amic mieu* » ; « *la lenga nòstra* » .

7. *a) Despartissètz **lo vòstre** ben e donatz-me çò que devi aver.*
   *b) Partiu **el vostre** bé i doneu-me el que m'escaigui.*
   'Partagez votre bien et donnez-moi ce qui me revient.'

8. *a) « O **mon filh** » diguèt lo paire, « coma voldràs tu ; siás un marrit e seràs castigat »*
   *b) « Ai, **fill meu** » , va dir el pare, « com vulguis ; ets dolent i seràs castigat »*
   'Ah, mon fils ! » dit le père, « comme tu voudras, tu es méchant et tu seras punis'

Le partitif comme dans l'exemple 9 en languedocien « *qu'an **de** pan e **de** vin, **d'**uòus e **de** formatge* » n'existe pas en catalan : « *que tenen pa i vi, ous i formatge* » .

9. *a) « Enlà, l'ostal del paire es plen de vailets qu'an **de** pan e **de** vin, **d'**uòus e **de** formatge tant que vòlon. »*
   b) « Allà a-baix, la casa del meu pare és plena de mossos que tenen pa i vi, ous i formatge, tant com en volen »
   '« Là-bas, la maison du père est pleine de serviteurs qui ont du pain, du vin, des œufs et du formage autant qu'ils en veulent. »'

Pour conclure, le Tableau 4 synthétise ces différences ci-dessous.

	Catalan standard	**Occitan**
Place de la négation simple	Avant le verbe	Après le verbe
Anar+verbe	Prétérit	Futur proche
Place des clitiques pour les verbes à l'infinitif	Après le verbe	Avant le verbe
Possessifs	Composés	Simples ou composés
Partitif	Non	Oui

Tableau 4. Synthèse des différences entre le catalan standard et l'occitan languedocien

Néanmoins, nous comptons sur tout ce qui rapproche les deux langues (pas de pronom clitique sujet, place de l'adjectif préférée derrière le nom, …) et posons l'hypothèse que la probabilité des séquences d'étiquettes grammaticales sera suffisamment similaire pour améliorer nos résultats en utilisant une ressource catalane pour entraîner un analyseur morphosyntaxique de l'occitan languedocien avec Talismane que nous présentons dans la section suivante.

3 Talismane

3.1 Fonctionnement

Dans cette étude, nous avons entrainé l'analyseur morphosyntaxique Talismane (Urieli, 2013), distribué sous une licence libre[4], sur des corpus d'entraînement occitans et catalans. Talismane a déjà été appliqué à l'anglais et au français, avec une exactitude de 97 % (Urieli, 2014), et à l'occitan avec une exactitude de 89 % (Vergez-Couret et Urieli, 2014). Talismane permet d'intégrer un lexique à la fois sous forme de descripteurs et de règles. En tant que descripteur, le lexique nous permet de dire « si le mot X existe dans le lexique en tant que nom commun, alors il est plus probable qu'il soit réellement un nom commun ». Cette information est incorporée dans le modèle statistique pendant l'entraînement, avec d'autres descripteurs listés ci-dessous. En tant que règle, le lexique nous permet de contourner les choix du modèle statistique pendant l'analyse, soit en imposant ou en interdisant le choix d'une certaine étiquette. Par exemple, on peut définir la règle suivant laquelle « le mot X ne peut être étiqueté comme préposition que s'il est listé comme préposition dans le lexique ».
Nous avons effectué une recherche de plusieurs combinaisons traditionnellement utilisées en apprentissage automatique et sélectionné la meilleure configuration pour l'occitan avec un classifieur SVM linéaire avec $\varepsilon = 0{,}1$ et $C = 0{,}5$.

3.1.1 Descripteurs

Nous utilisons le même jeu de descripteurs pour l'occitan que pour le français et l'anglais. Ces descripteurs ont été choisis en premier lieu en suivant l'intuition qu'ils indiqueront des distributions particulières d'étiquettes. Nous présentons ci-après ceux qui ont été validés et retenus après une évaluation empirique. Pour analyser un token T_i, on examine les tokens en position T_{i-2}, T_{i-1}, T_i, T_{i+1}, T_{i+2}. Les descripteurs de base comprennent pour chacun de ces tokens : **W** la forme lexicale exacte, **P** l'étiquette attribuée au token T_j (si $j < i$) ou les étiquettes trouvées dans le lexique (si $j \geq i$), **L** le lemme trouvé dans le lexique pour chaque étiquette donnée, **U** si le token est inconnu dans le lexique, $\mathbf{Sfx}_n$ les *n* dernières lettres de la forme (*n* de 2 à 5), **1st** si le token est le premier de la phrase, **Last** si le token est le dernier de la phrase. Ces briques de base sont aussi combinées en bigrammes et trigrammes. Ainsi, par exemple, $\mathbf{P}_{i-1}$ estime la distribution des probabilités pour l'étiquette du token actuel étant donné l'étiquette attribué au token précédent. $\mathbf{P}_{i-2}\mathbf{P}_{i-1}$ estime la distribution des probabilités pour cette même étiquette étant donné les étiquettes attribuées aux deux tokens précédents.

3.1.2 Règles

Les règles suivantes ont été définies autour des étiquettes des classes fermées (i.e. des catégories fonctionnelles non productives) et des classes ouvertes (i.e. des catégories lexicales productives).

- Classes fermées : l'analyseur peut attribuer une étiquette de classe fermée (e.g. prépositions, conjonctions, pronoms, …) uniquement si le token actuel est listé sous cette étiquette dans le lexique. Cette règle nous empêche, par exemple, d'inventer de nouvelles prépositions.

[4] http://redac.univ-tlse2.fr/talismane.html

- Classes ouvertes : l'analyseur ne peut pas attribuer une étiquette de classe ouverte (e.g. nom commun, adjectif, …) si le token en question est listé uniquement dans les lexiques des classes fermées. Cette règle nous empêche, par exemple, d'attribuer l'étiquette « nom commun » au token *lo* (« le » en français).

- Des règles basées sur les expressions régulières pour systématiquement attribuer les étiquettes des nombres cardinaux et de la ponctuation.

4 TA des langues peu dotées

4.1 Talismane pour l'occitan

Dans Vergez-Couret et Urieli (2014), un premier modèle de Talismane a été entraîné avec un corpus d'entraînement de 2 500 mots, un lexique de 225 386 entrées et plusieurs petits corpus d'évaluation d'environ 700 mots chacun avec pour objectif a) la création d'un premier modèle d'analyse morphosyntaxique pour l'occitan ; b) l'évaluation des performances de Talismane entraîné avec un corpus et un lexique du même dialecte, en l'occurrence languedocien, sur des corpus représentant une certaine variété dialectale ; c) l'évaluation du gain pour la création du modèle des corpus annotés *vs.* des lexiques afin de déterminer quel est l'effort le plus important pour la constitution de ressources des langues peu dotées.

4.2 Constitution des ressources : gain des corpus annotés et des lexiques

Les résultats de Vergez-Couret et Urieli (2014) tendent à montrer l'importance des lexiques, et notamment des lexiques des classes fermées (adverbes (quantifieurs, négatifs, exclamatifs, interrogatifs), déterminants, prépositions, pronoms, adjectifs possessifs). Cette conclusion va dans le sens de celle de Garrette et al. (2013) qui ont accompli une expérience où un temps limité de 4h était donné à des annotateurs pour annoter soit du corpus, soit des lexiques (construits à partir des mots les plus fréquents de corpus non annotés) pour deux langues peu dotées. Ils concluent que le gain le plus important est obtenu avec la création des lexiques.
Or s'il est toujours possible d'augmenter la quantité de corpus, cela est moins vrai pour les lexiques. Dans cet article, nous souhaitons donc nous focaliser sur le seul dialecte languedocien avec l'objectif d'obtenir le meilleur résultat possible en ne créant pas ou peu de nouvelles ressources annotées et donc en utilisant et en adaptant des ressources d'une langue étymologiquement proche.

4.3 Utilisation de ressources de langues étymologiquement proches

Dans le cas des langues peu dotées en ressources linguistiques et de TAL, des méthodes basées sur l'adaptation d'analyseurs morphosyntaxiques des langues très dotées, généralement étymologiquement proches, ont récemment été développées. Täckström et al. (2013) utilise une approche semi-supervisée basée sur un bitexte qui aligne une langue peu dotée et une langue très dotée et ont obtenu un gain significatif. Scherrer et Sagot (2013) utilisent une approche visant à identifier des cognats lexicaux entre une langue peu dotée et une langue très dotée étymologiquement proche pour récupérer la catégorie grammaticale du mot dans la langue très dotée et améliorer l'annotation de la langue peu dotée. Cette approche a l'avantage de ne nécessiter aucune ressource annotée pour la langue peu dotée. Dans le cas de l'occitan, nous ne disposons pas de bitextes pour l'occitan et une autre langue très dotée, ce qui élimine la première méthode. En revanche, nous disposons d'un lexique avec une assez bonne couverture (que nous présentons section 4.1) et des corpus annotés et déjà exploités dans Vergez-Couret et Urieli (2014), ce qui exclut finalement la deuxième méthode. Dans cet article, notre objectif est d'améliorer les résultats obtenus dans Vergez-Couret et Urieli (2014) en augmentant les corpus et en comparant les apports d'un petit corpus annoté en languedocien correspondant exactement au besoin attendu (qualité des annotations) et un grand corpus catalan (quantité des annotations) et en expérimentant plusieurs approches de transposition et de combinaison.

5 Ressources

Nous allons présenter dans cette section toutes les ressources qui ont été rassemblées, nécessaires à l'entraînement de Talismane et à la mise en œuvre de nos expériences.

5.1 Lexique et jeu d'étiquettes

La version du lexique que nous présentons ici est une version étendue de 50 000 formes (principalement de classes lexicales (adjectif, verbe, nom)) de la version utilisée dans Vergez-Couret & Urieli (2014). Ce lexique concerne uniquement la variante languedocienne de l'occitan. Il a principalement été construit avec une ressource disponible au format numérique : le dictionnaire Français/Occitan languedocien de C. Laus (2005). Les noms propres ont été extraits des lexiques Apertium (Armentano-Oller & Forcada, 2006). Des listes de formes fléchies ont été rassemblées à partir du conjugueur mis à disposition par le *Congrès permanent de la lenga occitana*. Enfin, un script a permis de générer les formes fléchies des adjectifs, des noms et des participes passés ainsi que les formes élidées et contractées des prépositions et des déterminants. Le nombre d'entrées de chaque catégorie principale est disponible dans le Tableau 5.

Etiquette	Description	Taille
A	Adjectif (général)	29656
A$	Adjectif (possessif)	85
Adv	Adverbe (général)	762
Adv$	Adverbe (négatif, quantifieur, exclamatif et interrogatif)	58
Cc	Conjonction de coordination	8
Cs	Conjonction de subordination	150
Det	Déterminant	127
Card	Cardinal	42
Cli	Pronom clitique	17
CliRef	Pronom réfléchi	17
Inj	Interjection	130
Nc	Nom commun	53449
Np	Nom propre	4609
Pct	Ponctuation	15
Pp	Participe présent	4554
Pr	Préposition	521
Prel	Pronom relatif	37
Pro	Pronom non clitique	81
Ps	Participe passé	18089
PrepDet	Préposition et déterminant amalgamé	499
Vc	Verbe conjugué	160549
Vi	Verbe à l'infinitif	5822
Z	Consonnes de liaison	3
	Total	279280

Tableau 5. Nombre de formes fléchies du lexique

5.2 Corpus d'entraînement

Pour cette étude, deux corpus d'entraînement seront utilisés : un corpus de petite taille en languedocien que nous avons annoté et un corpus de grande taille en catalan, le corpus Ancora (Taulé et al., 2008).

Le corpus languedocien (Occitan-Train) est composé d'un extrait d'une œuvre littéraire *E la barta floriguèt* d'Enric Molin et d'un article de wikipedia occitan. Ce corpus contient 3000 formes annotées manuellement avec le lemme, la catégorie grammaticale et des informations morphosyntaxiques (genre, nombre, personne, temps et aspect). Les 1000 premières formes ont été séparément annotées par trois annotateurs qui ont ensuite confronté leurs annotations pour décider d'une annotation commune et répertorier les décisions dans un manuel d'annotation. Puis 1500 mots ont été annotés par deux annotateurs qui ont également confronté leurs annotations pour décider d'une annotation commune et améliorer le manuel d'annotation. Enfin, les 500 derniers mots n'ont été annotés que par un seul annotateur.

Le corpus catalan (Catalan-Train) est une adaptation avec notre jeu d'étiquettes du corpus Ancora. Il contient 500 000 formes principalement de textes journalistiques. Le corpus Ancora est un corpus annoté à plusieurs niveaux (morphosyntaxique, syntaxique et sémantique) et contient en particulier toutes les annotations qui nous intéressent : le lemme, la catégorie grammaticale et les informations morphosyntaxiques.

5.3 Corpus d'évaluation

Pour cette étude, nous avons un corpus d'évaluation contenant des extraits de deux œuvres littéraires *Los crocants de Roergue* de Ferran Delèris (700 formes) et *Dels camins bartassièrs* de Marceu Esquieu (500). Le premier corpus a été annoté par deux annotateurs qui se sont mis d'accord sur une annotation commune après avoir confronté leurs annotations tandis que le second n'a été annoté que par un seul annotateur.

Bien que les informations morphosyntaxiques soient disponibles dans tous nos corpus (ex. genre, nombre), elles ne sont employées ni pour l'entraînement, ni pour l'évaluation. La tâche d'étiquetage pour la catégorie principale est déjà difficile pour une langue peu dotée et une fois que l'étiquette principale déterminée, les informations morphosyntaxiques sont souvent disponibles dans le lexique.

5.4 Adaptation des ressources

Une des difficultés rencontrées avec des données extraites de plusieurs sources et de plusieurs langues fut le manque de consistance des annotations entre les corpus d'entraînement et le corpus d'évaluation, les jeux d'étiquettes et les règles d'annotation pouvant être légèrement différents d'un corpus à l'autre.

De ce fait, nous avons dû procéder à des harmonisations. Une première harmonisation a été de choisir une étiquette plus générale qui couvre les annotations des deux corpus. Nous avons ainsi transformé les étiquettes CLI (pronom clitique), CLIREF (pronom réfléchi) en PRO (pronom) dans les corpus et les lexiques étant donné que les deux distinctions précédentes n'étaient pas annotées dans le corpus Ancora. Ensuite, nous avons ajouté une distinction au corpus catalan entre les adverbes de classes ouvertes (par exemple les adverbes en *–ment*) et les adverbes de classes fermées (négatifs, quantifieurs, …) sur la base d'une liste.

En plus des différences de formation et d'emploi des possessifs que nous avons brièvement abordés dans la section 1.3.3, les normes d'annotation des possessifs varient entre le corpus d'entraînement catalan et le corpus d'entraînement languedocien. Les formes possessives complexes ont été segmentées dans le corpus languedocien (« la sia » est annotée en deux formes : « la » est annotée comme déterminant défini et « sia » est annoté comme adjectif possessif) et pas dans le corpus catalan (« la_seva » est dans son ensemble annotée comme déterminant possessif). Dans ce cas, harmoniser nécessiterait une intervention au niveau de la segmentation. Nous n'avons pas effectué cette harmonisation et savons que l'annotation des adjectifs possessifs risque de ne pas être faite selon nos besoins.

5.5 Transposition du corpus catalan

Pour cette étude, nous avons adopté une méthode visant à transposer les mots plus fréquents du corpus catalan en occitan. Ce type de méthode est généralement employée pour transposer les mots les plus fréquents dans les textes d'une langue peu dotée vers leurs équivalents dans une langue très dotée étymologiquement proche (Bernhard et Ligozat, 2013 ; Vergez-Couret, 2013). Puis, l'analyseur morphosyntaxique de la langue très dotée est utilisé pour annoter la langue peu dotée. Dans cette expérience, nous faisons la transposition de la langue très dotée vers la langue peu dotée pour améliorer la couverture du lexique lors de l'entraînement et dans le but final d'améliorer les résultats lors de l'analyse. L'avantage d'effectuer la transposition dans ce sens est de pouvoir s'appuyer sur les catégories morphosyntaxiques pour transposer correctement les homographes. Nous avons construit un lexique bilingue languedocien/catalan en prenant les 250 mots catalans les plus fréquents du corpus Ancora que nous avons manuellement traduit en occitan languedocien. Nous obtenons un lexique de 150 paires de conversion (100 formes étant identiques entre les deux langues), principalement des mots grammaticaux. Les paires contenues dans la liste, une fois automatiquement transposées du catalan vers l'occitan dans Catalan-Train couvrent 93 243 changements, soit environ 19% des formes du corpus.

Les tailles des corpus et la couverture des lexiques sont synthétisés dans les tableaux ci-dessous :

Corpus	*Occitan-Train*	*Catalan-Train*	*Catalan-Train Transposé*	*Occitan-Eval*
Taille	2 840	488 389	488 389	1 214
Taille totale (sans la ponctuation) *% de formes inconnues dans le lexique*	2 368 *2,4 %*	435 814 *39,4 %*	435 814 *30,7 %*	1 018 *9,3 %*
Classes ouvertes *% de formes inconnues dans le lexique*	1 282 *4,3 %*	225 487 *62,8 %*	225 487 *51,8 %*	543 *16,9 %*
Classes fermées *% de formes inconnues dans le lexique*	1 086 *0,1 %*	205 958 *12,5 %*	205 958 *6,0 %*	475 *0,6 %*

Tableau 6. Corpus d'entraînement et d'évaluation

6 Expériences et résultats

Les ressources ont été rassemblées pour répondre aux questions suivantes :
1) Faut-il favoriser des annotations de qualité (corpus languedocien) ou des annotations en quantité (corpus catalan) ?
2) Est-il utile de faire « ressembler » le texte catalan à de l'occitan ?
3) Peut-on améliorer la qualité des résultats en combinant les deux corpus et sous quelles conditions ?

6.1 Entraînement avec un petit languedocien et un gros catalan

La première expérience vise tout simplement à entraîner Talismane avec nos corpus d'entraînement languedocien et catalan séparément : Occitan-Train (3 000 formes) ; un extrait de 3 000 formes de Catalan-Train et Catalan-Train.

Corpus d'entraînement	Occitan-Train	Catalan-Train (3 000 formes)	Catalan-Train
Exactitude	89,04	86,00	90,11

Tableau 7. Résultats

A taille égale, les annotations en languedocien permettent sans grande surprise une amélioration significative (*p*-valeur < 0,005, test de McNemar) par rapport aux annotations en catalan. Utiliser un gros corpus catalan (90,5 %) permet une petite amélioration non significative (*p*-valeur >0,1, test de McNemar) comparé au petit corpus languedocien (89 %). Nous avons souhaité voir s'il est encore possible de faire une amélioration significative de la baseline (Occitan-Train) a) en modifiant le corpus catalan (transposition en occitan des mots les plus fréquents) et b) en combinant les deux corpus languedocien et catalan.

6.2 Transposition des mots les plus fréquents

Même si les gains obtenus avec le corpus catalan ne sont pas significatifs, nous avons constaté l'impact très positif du lexique lors de l'entraînement : en effet, sans lexique, les scores baissent à 53,77%. Selon nous, le lexique fonctionne dans ce cas grâce au grand nombre de cognats (homographes entre le catalan et l'occitan qui partagent la même étiquette morphosyntaxique). Ainsi, lors de l'entraînement, Talismane « apprend » que si un mot se trouve dans le lexique, il a une forte probabilité d'être employé dans le corpus d'analyse avec la même étiquette. L'idée sous-tendant la présente expérience revient alors à faire « ressembler » encore plus le corpus catalan à de l'occitan (et à améliorer la couverture du lexique) en transposant en occitan les mots les plus fréquents du corpus catalan. 150 couples ont été automatiquement transposés du catalan vers l'occitan (sans vérification manuelle), concernant environ 19 % du corpus (cf. section 4.5) permettant de passer de 39,4 à 30,7 % de formes inconnues dans le lexique (cf. Tableau 6).

Corpus d'entraînement	Catalan-Train	Catalan-Train (transposé)
Exactitude	90,69	91,10

Tableau 8. Résultats avec transposition des mots les plus fréquents

La transposition permet une petite amélioration significative (*p*-valeur < 0,05, test de McNemar) sur l'intégralité du corpus Catalan-Train. Mais surtout, cela permet pour la première fois une amélioration significative de +2,06% par rapport à la baseline (*p*-valeur < 0,01, test de McNemar). Il serait intéressant d'augmenter le nombre de transposition des

mots les plus fréquents mais également des classes lexicales, par exemple en utilisant des méthodes d'appariement de cognats (cf. conclusion). La prochaine expérience que nous présentons vise à combiner les deux corpus d'entraînement catalan (version transposée) et languedocien.

6.3 Combinaison du corpus catalan et du corpus occitan languedocien (avec pondération)

Nous avons créé un corpus d'entraînement unique afin de joindre qualité et quantité à partir des deux corpus catalan et occitan languedocien. Puis nous avons testé plusieurs configurations faisant varier le poids du corpus occitan languedocien (en dupliquant *x* fois le corpus). Les 3 000 formes annotées en languedocien se trouvent noyées parmi les 500 000 formes annotées en catalan et la combinaison des deux corpus ne permet pas de gain significatif. Mais donner plus de poids au corpus languedocien permet d'améliorer sensiblement les résultats jusqu'à notre meilleur résultat de 92,26 avec un poids de 200 attribué au corpus languedocien, donc un gain de +1,16% par rapport au corpus catalan seul transposé (*p*-valeur = 0,0007, test binomial) et +3,22% par rapport à la baseline (*p*-valeur < 0,005, test de McNemar)..

Corpus d'entraînement	Catalan-Train + Occitan-Train ×1 (≈ 3 000)	Catalan-Train + Occitan-Train ×25 (≈ 75 000)	Catalan-Train + Occitan-Train ×50 (≈ 150 000)	Catalan-Train + Occitan-Train ×100 (≈ 300 000)	Catalan-Train + Occitan-Train ×200 (≈ 600 000)
Exactitude	91,26	91,68	92,17	92,00	92,26

Tableau 9. Résultats selon le poids attribué au corpus languedocien

7 Conclusion

Dans cet article, nous souhaitions montrer qu'il est possible d'obtenir de meilleurs résultats que Vergez-Couret et Urieli (2014) pour l'analyse morphosyntaxique de l'occitan en employant une ressource d'une langue bien dotée et étymologiquement proche, le catalan. En effet, nous avons amélioré les résultats de 3,22 % en atteignant 92,26 % en combinant une ressource catalane de grande taille, Ancora, et une petite ressource annotée en occitan languedocien et en attribuant un poids plus important au corpus languedocien. Le résultat le plus intéressant de notre point de vue est d'avoir montré par ce biais que même un petit corpus (1/200ème du corpus catalan), plus proche de la variante à évaluer, peut améliorer les scores de façon significative, ce qui est très prometteur pour l'analyse inter-variante.

Par ailleurs, nous avons testé des méthodes visant à « faire ressembler » superficiellement le corpus catalan à l'occitan en remplaçant les 250 formes lexicales les plus fréquentes par leur traduction. Les résultats sont encourageants pour 19 % de transformations sur le corpus catalan, essentiellement des catégories grammaticales. Il serait donc intéressant d'étendre la couverture des transformations aux catégories lexicales pour voir si plus de régularités de formes permettent un gain significatif. Pour ce faire, nous pensons nous inspirer des méthodes de similarité lexicale présentées dans Scherrer et Sagot (2013).

Pour cette expérience, nous avons peu évalué l'effet du genre dans les corpus d'entraînement et lors de l'évaluation, bien qu'il a été montré que cela peut avoir un effet marqué (Candito et Seddah, 2012). Notre objectif est en premier lieu d'annoter les textes littéraires de la base BaTelòc (Bras et Thomas, 2011) (dont est extrait le corpus d'évaluation), mais dès que BaTelòc va s'ouvrir à d'autres types de textes, il sera intéressant d'augmenter la taille et la variété de nos ressources d'entraînement et d'évaluation. Afin de mieux comprendre les processus d'analyse et améliorer les résultats, on serait tenté d'analyser les erreurs dans nos corpus d'évaluation. Mais cela risquerait d'induire une sur-adéquation du modèle à moins de les passer en ressources d'entraînement. Ceci définit une méthodologie de travail dans laquelle les données d'évaluation de chaque étude deviennent des ressources de développement pour l'étude suivante. A terme, un gros occitan viendra-t-il renforcer l'amitié catalano-occitane ? *A la fortuna !*

Remerciements

Nous remercions les trois relecteurs anonymes du comité scientifique de TALaRE ainsi qu'Estel Llansana et Nabil Hathout pour leurs conseils qui ont permis d'améliorer la qualité de l'article. Les travaux décrits dans cet article ont bénéficié du soutien de l'ANR (projet RESTAURE - convention ANR-14-CE24-0003-01).

Références

ALMARCHA PARIS M., BAYLAC FERRER, A. (2007). La langue des Pays Catalans. *Langues et Cité :bulletin de l'observation des pratiques linguistiques,* 21, 2.

ARMENTANO-OLLER, C., FORCADA M.-L. (2006). Open-source machine translation between small languages: Catalan and Aranese Occitan. *In Strategies for developing machine translation for minority languages (5th SALTMIL workshop on Minority Languages organized in conjunction with LREC 2006)*, pp. 51-54.

BEC P. (1995). *La langue occitane*. Number 1059. Paris : Que sais-je ?

BERNHARD, D., LIGOZAT A.-L. (2013). Es esch fàscht wie Ditsch, oder net? Étiquetage morpho-syntaxique de l'alsacien en passant par l'allemand. *In Actes de TALARE 2013 : Traitement Automatique des Langues Régionales de France et d'Europe*, pp. 209-220.

BOLEDA G., CUADROS, M., ESPANA-BONET C., MELERO, M., PADRO, L., QUIXAL, M. RODIRGUEZ, C. (2007). Primera Jornada del Procesamiento Computacional del Catalán. *Processamiento del Lenguaje Natural, núm,* 43, 387-388.

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.

CANDITO M. et SEDDAH D. (2012). Le corpus sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. *In TALN 2012-19e conférence sur le Traitement Automatique des Langues Naturelles.*

GARETTE, D., MIELENS J., BALDRIDGE J. (2013). Real-word semi-supervised learning of pos-taggers for low-resource languages. *In Actes dela conférence de l'Association for computational linguistics (ACL)*, pp. 583-592.

LAUS C. (2005). *Dictionnaire Français-Occitan*. Castres : IEO del Tarn.

SCHERRER Y., SAGOT B. (2013). Lexicon induction and part-of-speech tagging of non-resourced and tools for closely related languages and language variants. Actes de *RANLP Workshop on Adaptation of language resources and tools for closely related languages and language variants*, 30-39.

SERRA SERRA J. (2007). Le catalan sur internet. *Langues et Cité :bulletin de l'observation des pratiques linguistiques,* 21, 11.

SIBILLE J. (1996). Lo gascon dialecte occitan o lenga a part entièra : Es que la question a un sens ? Elements de responsa a las teorias de Jan Lafita. *Estudis Occitans,* 20, 38-40.

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

TACKSTROM O., DAS D., PETROV S., MCDONALD R., NIVRE J. (2013). Token and type constraints for cross-lingual part-of-speech tagging. Actes de *Transactions of the Association for Computational Linguistics*, 1-12.

TAULE M., MARTI M.A., RECASENS M. (2008). Ancora: Multilevel Annotated Corpora for Catalan and Spanish. Actes de *6th International Conference on Language Resources and Evaluation*, 96-101.

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

URIELI A. (2014). Améliorer l'étiquetage de « que » par les descripteurs ciblés et les règles. *In Actes de la 21ème conférence sur le Traitement Automatique des Langues Naturelles;* 56-66

VERGEZ-COURET M. (2013). Tagging Occitan using French and Castillan Tree Tagger. *In Actes de Less Resources Languages, new technologies, new challenges and opportunities workshop in conjunction with the 6th Language & Technology Conference;* 56-66

VERGEZ-COURET M., URIELI A. (2014). POS-tagging different varieties of Occitan with single-dialect ressources. Actes de *The First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, Dublin, 21-29.