Le traitement automatique des langues peu dotées

Le cas de l'occitan

Marianne Vergez-Couret - Assaf Urieli

UE TAL - Master 2

CLLE-ERSS - UMR 5263
Université de Toulouse

and

Joliciel Informatique
Foix, France

20 octobre 2014

1 Occitan

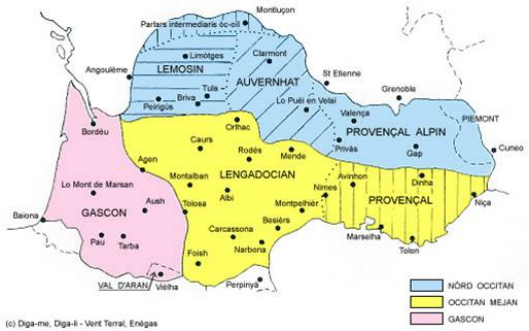
Contexte

Traitement automatique des langues peu dotées

Particularités de l'occitan

Situation et contexte politique

- Langue romane
- Organisée en dialectes
- Couvrant 8 régions administratives
- Pas de statut officiel en France



(Extrait de *Diga-me, diag-li*, Vent Terral, Enègas)

Contexte sociolinguistique

- Nombre de locuteurs : environ 500 000 sur une population générale de 15 millions (3%)
- Enquête en Midi-Pyrénées : Natifs ou Bilingues : 4% ; Niveau intermédiaire : 14% ; Locuteurs passifs : 32%
- Enseignement
- Présence dans les médias (presse, web, radio, télé)
- Réseaux associatifs

Développement numérique de l'occitan

- Principaux acteurs (Formation Diagnostic numeric occitan)
  - Lo congrés permanent de la lenga occitana (dictionnaire en ligne, conjugueur, corpus interrogeable via un concordancier)
  - CIRDOC (Occitanica, médiathèque numérique occitane)
  - Les médias (presse, radio, télé)
- Besoins
  - OCR (reconnaissance d'écriture (manuscrit et tapuscrit))
  - Traduction automatique
  - Synthèse vocale

Un véritable défi pour le TAL

- Faible rentabilité financière de l'informatisation qui ne compense pas les coups de développement considérables (humains et financiers)
- Systèmes robustes pour gérer le manque de ressources et la variation

... et pour l'occitan

- Assurer la collecte des données, utiliser des formats normalisés pour diffusion, pérennité, réutilisabilité
- Crucial pour la sauvegarde, la transmission et l'enseignement de l'occitan
- Enrichir les recherches en sciences humaines et sociales (linguistique, sociologie, littérature, histoire)

Langue écrite

- 1000 ans de littérature
- Pas de standardisation pour la langue dans son ensemble mais émergence de formes plus ou moins standardisées pour chaque dialecte
- Plusieurs systèmes graphiques :
  - Moyen-Age : graphie des troubadours
  - 19ème siècle : graphies inspirées de la graphie française
  - 20ème siècle : graphie classique

Langue romane

Français	Italiano	Castillano	Portugues	Català	Occitan (Lengadoc
mouche	mosca	mosca	mosca	mosca	mosca
amie	amica	amiga	amiga	amiga	amiga
amour	amore	amor	amor	amor	amor
chèvre	capra	cabra	cabra	cabra	cabra
château	castello	castillo	castelo	castell	castèl
table	tavolo	mesa	mesa	taula	taula

Variétés dialectales

Lengadocian	Auvernhat	Gascon	Lemosin	Provençau	Vivaroaup
mosca	moscha	mosca	moscha	mosca	moissa
amiga	amia	amiga	amiga	mia	amia
amor	amor	amor	amor	amor	amor
cabra	chabra	craba	chabra	cabra	chabra
castèl	chastèl	castèth	chasteu	castèu	chasteu
taula	tala/taula	taula	taula	taula	taula
nuèch/nuèit /nuòch	neut/nueit	nèit/nuèit /nueit/neit /net	nuech	nuech	nuech

Variantes graphiques

gniu, gnoch, gnué, nè, nèch, nèi, nèit, nèt, nèyt, nét, néyt, neit, net, neu, neuit, neut, ney, neyt, niè, nièch, nieu, nio, nioch, niu, niue, niuech, niuit, noéyt, not, nou, noueit, nuè, nuèch, nue, nuech, nueit, nuet, nueyt, nuyt

❷ BaTelÒc

Motivations

- Besoin de ressources pour travailler sur l'occitan
- Se constituer son propre corpus
- Construire une base textuelle pour l'occitan

Objectifs

- Etape 1
  - Rassembler des oeuvres écrites de différents genres, des époques modernes et contemporaines
  - Accueillant toute la variation (dialectale et graphique) possible
- Etape 2
  - Création d'outils pour la sélection des corpus et l'exploration des textes (concordancier)
- Etape 3
  - Enrichir d'annotations linguistiques

Objectifs

- Etape 1
  - Rassembler des oeuvres écrites de différents genres, des époques modernes et contemporaines
  - Accueillant toute la variation (dialectale et graphique) possible
- Etape 2
  - Création d'outils pour la sélection des corpus et l'exploration des textes (concordancier)
- Etape 3
  - Enrichir d'annotations linguistiques

Objectifs

- Etape 1
  - Rassembler des oeuvres écrites de différents genres, des époques modernes et contemporaines
  - Accueillant toute la variation (dialectale et graphique) possible
- Etape 2
  - Création d'outils pour la sélection des corpus et l'exploration des textes (concordancier)
- Etape 3
  - Enrichir d'annotations linguistiques

Stratégies pour la constitution de la base

- Commencer par les textes déjà numérisés, puis remonter dans le temps (scan et OCR)
- Codage XML (réutilisabilité)
- Bâtir des partenariats avec le milieu occitan (éditeurs, bibliothèques virtuelles...)

Présentation

- Petite base (environ 60 textes, 2 millions de mots)
- Genres : roman, conte, poésie, essai, mémoires
- Outils pour construire un corpus de travail
- Concordancier
- Sortie prévue printemps 2015

Stratégies pour l'enrichissement avec des annotations linguistiques

- Commencer par l'analyse morphosyntaxique
- Annotation d'un sous-ensemble de la base (cohérent d'un point de vue dialectal)
- Etape 1 : Avec des outils existants (cf. Apertium)
- Etape 2 : Avec une plateforme générique d'entraînement par apprentissage supervisé

❸ Langues peu dotées dans RESTAURE

Stratégie globale

- Travailler avec les autres langues peu dotées
- RESTAURE : Ressources informatisées et Traitement AUtomatique des langues REgionales de France
- Alsacien, Occitan, Picard
- Mutualiser les outils
- Mutualiser les expériences

Objectifs

- Acquisition et normalisation de ressources (corpus, lexiques, dictionnaires).
  - Ressources représentant un ensemble de variétés le plus large possible (genre, dialecte, graphie)
  - Scan de textes imprimés et océrisation
  - Encoder les ressources en suivant les formats standards
- Développement d'outils de TAL
  - OCR,
  - Segmentation des textes en phrases et en mots,
  - Analyse morphosyntaxique,
  - Lemmatisation et normalisation,
  - Reconnaissance des entités nommées
  - Désambiguïsation du sens des mots pour la traduction.

Méthodes

- Adapter les outils des langues proches
- Adapter les ressources des langues proches
- Même chose pour les dialectes

- En utilisant des méthodes par apprentissage supervisé
- Laboratoire expérimental pour les langues peu dotées
- Où mettre l'éffort ? (constitution et gestion des ressources)

4 Nos travaux

OCR

Analyse morphosyntaxique

OCR - Principes

- Jochre : Java Optical CHaracter REcognition
  - Logiciel libre développé par Assaf Urieli
- Apprentissage automatique supervisé
  - Annotation d'un corpus d'entraînement avec JochreWeb
- 3 étapes d'analyse :
  - segmentation des images en paragraphes, lignes, groupes et formes
  - reconnaissance des lettres
  - correction des mots à l'aide du lexique (re-ranking)

OCR - Principes

- Jochre : Java Optical CHaracter REcognition
  - Logiciel libre développé par Assaf Urieli
- Apprentissage automatique supervisé
  - Annotation d'un corpus d'entraînement avec JochreWeb
- 3 étapes d'analyse :
  - segmentation des images en paragraphes, lignes, groupes et formes
  - reconnaissance des lettres
  - correction des mots à l'aide du lexique (re-ranking)

OCR - Principes

- Jochre : Java Optical CHaracter REcognition
  - Logiciel libre développé par Assaf Urieli
- Apprentissage automatique supervisé
  - Annotation d'un corpus d'entraînement avec JochreWeb
- 3 étapes d'analyse :
  - segmentation des images en paragraphes, lignes, groupes et formes
  - reconnaissance des lettres
  - correction des mots à l'aide du lexique (re-ranking)

OCR - Principes

- Jochre : Java Optical CHaracter REcognition
  - Logiciel libre développé par Assaf Urieli
- Apprentissage automatique supervisé
  - Annotation d'un corpus d'entraînement avec JochreWeb
- 3 étapes d'analyse :
  - segmentation des images en paragraphes, lignes, groupes et formes
  - reconnaissance des lettres
  - correction des mots à l'aide du lexique (re-ranking)

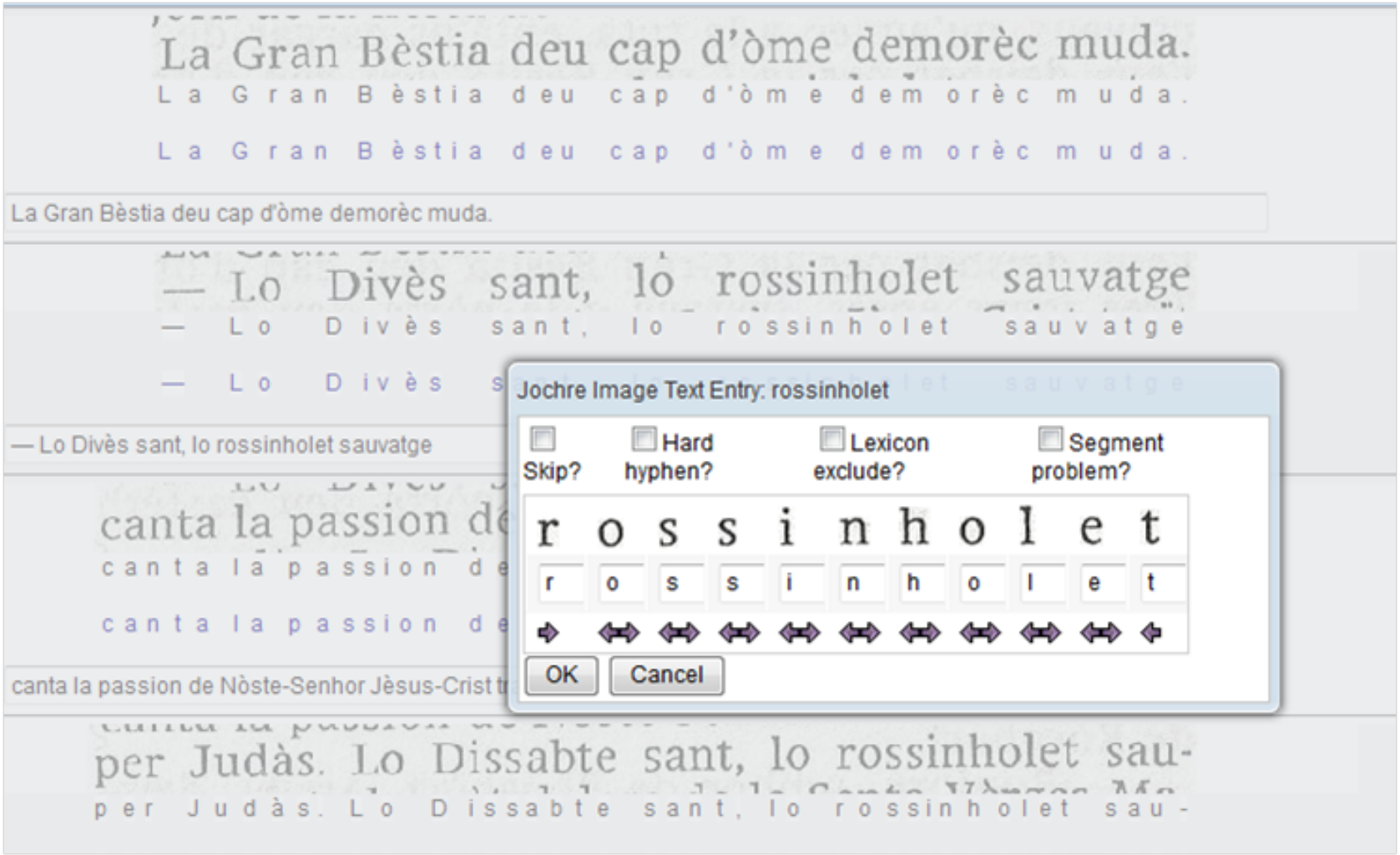
OCR - Principes

- Jochre : Java Optical CHaracter REcognition
  - Logiciel libre développé par Assaf Urieli
- Apprentissage automatique supervisé
  - Annotation d'un corpus d'entraînement avec JochreWeb
- 3 étapes d'analyse :
  - segmentation des images en paragraphes, lignes, groupes et formes
  - reconnaissance des lettres
  - correction des mots à l'aide du lexique (re-ranking)

OCR - Principes

- Jochre : Java Optical CHaracter REcognition
  - Logiciel libre développé par Assaf Urieli
- Apprentissage automatique supervisé
  - Annotation d'un corpus d'entraînement avec JochreWeb
- 3 étapes d'analyse :
  - segmentation des images en paragraphes, lignes, groupes et formes
  - reconnaissance des lettres
  - correction des mots à l'aide du lexique (re-ranking)

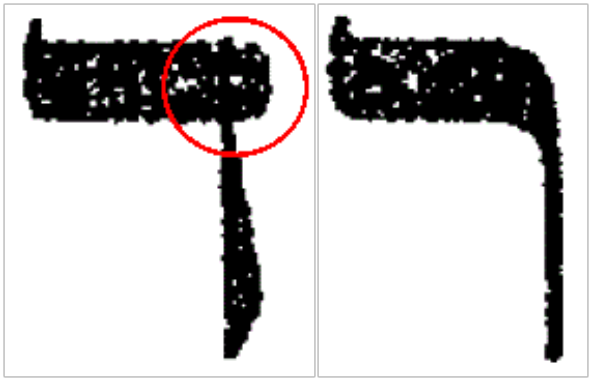
Annotation OCR avec JochreWeb



Analyse OCR Etape 1 - Segmentation

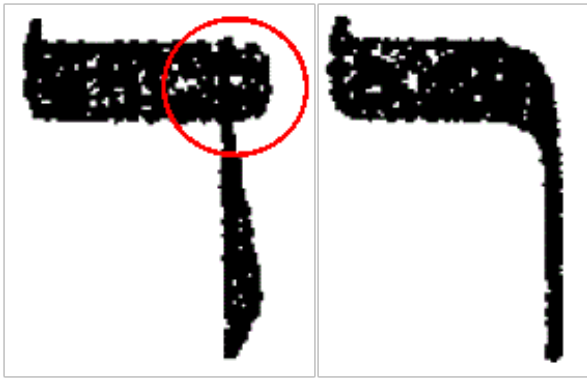
Analyse OCR Etape 2 - Reconnaissance des lettres

- Apprentissage automatique supervisé
- Annotation du corpus d'apprentissage sur JochreWeb
- Entraînement d'un modèle statistique par langue
- Descripteurs :
  - Descripteurs génériques (noirceur relative des cases d'une grille)
  - Descripteurs spécialisés (pour distinguer deux lettres proches)



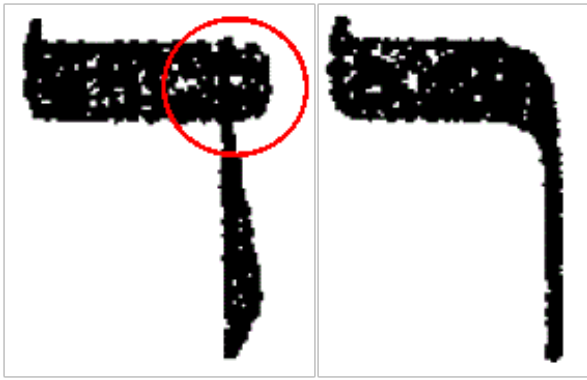
Analyse OCR Etape 2 - Reconnaissance des lettres

- Apprentissage automatique supervisé
- Annotation du corpus d'apprentissage sur JochreWeb
- Entraînement d'un modèle statistique par langue
- Descripteurs :
  - Descripteurs génériques (noirceur relative des cases d'une grille)
  - Descripteurs spécialisés (pour distinguer deux lettres proches)



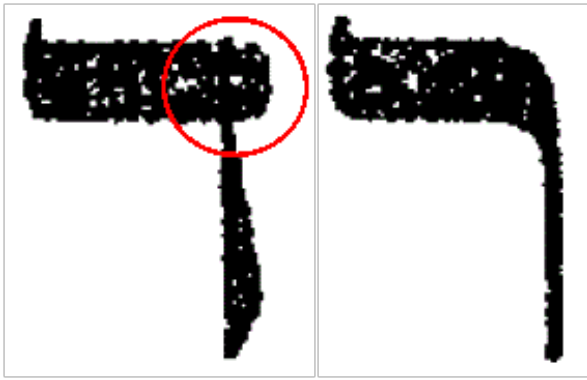
Analyse OCR Etape 2 - Reconnaissance des lettres

- Apprentissage automatique supervisé
- Annotation du corpus d'apprentissage sur JochreWeb
- Entraînement d'un modèle statistique par langue
- Descripteurs :
  - Descripteurs génériques (noirceur relative des cases d'une grille)
  - Descripteurs spécialisés (pour distinguer deux lettres proches)



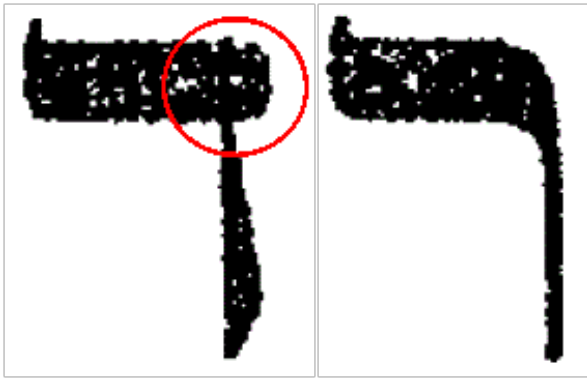
Analyse OCR Etape 2 - Reconnaissance des lettres

- Apprentissage automatique supervisé
- Annotation du corpus d'apprentissage sur JochreWeb
- Entraînement d'un modèle statistique par langue
- Descripteurs :
  - Descripteurs génériques (noirceur relative des cases d'une grille)
  - Descripteurs spécialisés (pour distinguer deux lettres proches)



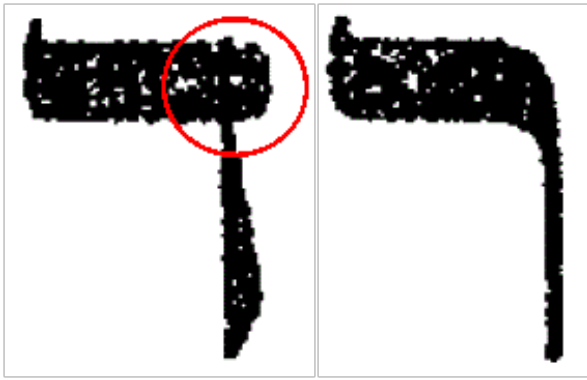
Analyse OCR Etape 2 - Reconnaissance des lettres

- Apprentissage automatique supervisé
- Annotation du corpus d'apprentissage sur JochreWeb
- Entraînement d'un modèle statistique par langue
- Descripteurs :
  - Descripteurs génériques (noirceur relative des cases d'une grille)
  - Descripteurs spécialisés (pour distinguer deux lettres proches)



Analyse OCR Etape 2 - Reconnaissance des lettres

- Apprentissage automatique supervisé
- Annotation du corpus d'apprentissage sur JochreWeb
- Entraînement d'un modèle statistique par langue
- Descripteurs :
  - Descripteurs génériques (noirceur relative des cases d'une grille)
  - Descripteurs spécialisés (pour distinguer deux lettres proches)



Analyse OCR Etape 3 - Correction (Reranking)

- Recherche par faisceau : les *n* analyses les plus probables
- Utilisation du lexique pour « reranking »

	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 %

- Possibilité de prendre en compte la fréquence

Analyse OCR Etape 3 - Correction (Reranking)

- Recherche par faisceau : les *n* analyses les plus probables
- Utilisation du lexique pour « reranking »

	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 %

- Possibilité de prendre en compte la fréquence

Analyse OCR Etape 3 - Correction (Reranking)

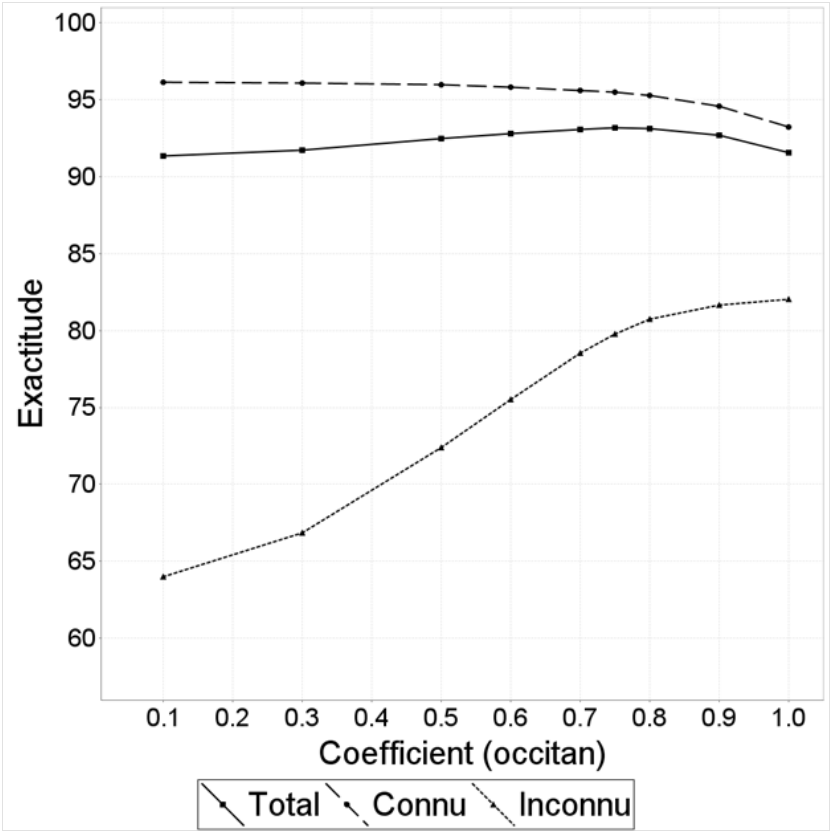
- Recherche par faisceau : les *n* analyses les plus probables
- Utilisation du lexique pour ≪ reranking ≫

	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 %

- Possibilité de prendre en compte la fréquence

OCR - paramètres re-ranking

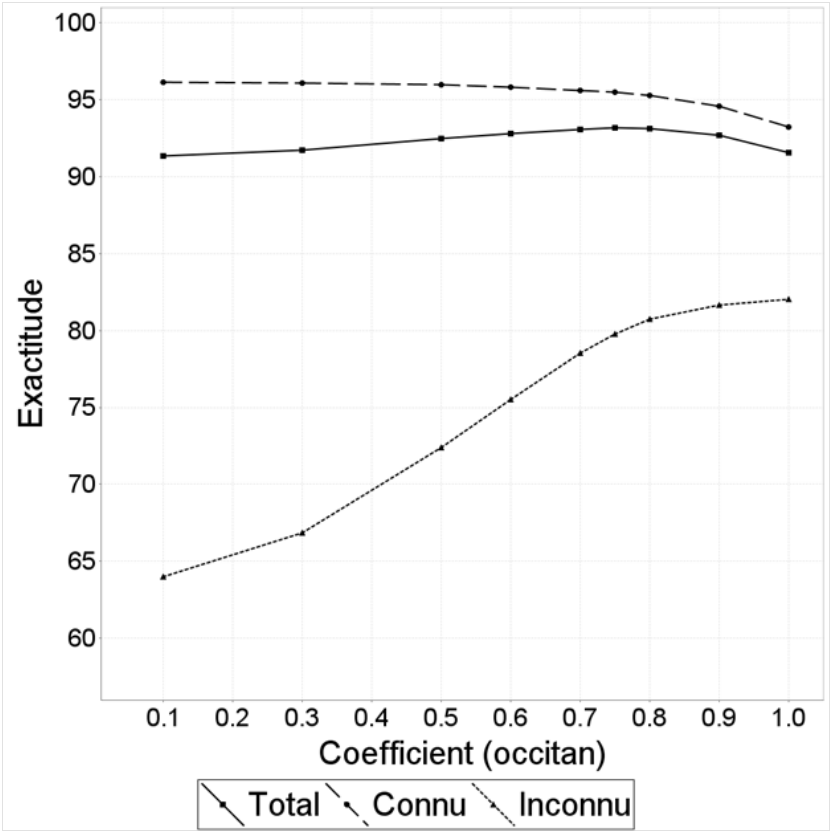
- Largeur de faisceau : 1, 2, 5, **10**, 20
- Coefficient de réduction pour mots inconnus : **0,75**



- Prise en compte de la fréquence ? **Non**

OCR - paramètres re-ranking

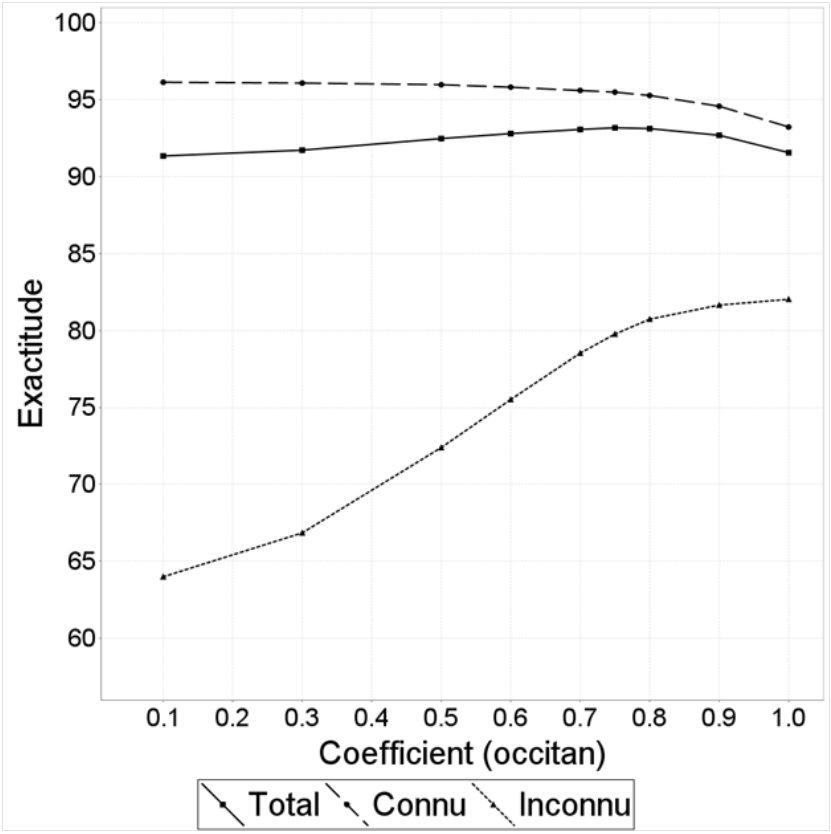
- Largeur de faisceau : 1, 2, 5, **10**, 20
- Coefficient de réduction pour mots inconnus : **0,75**



- Prise en compte de la fréquence ? **Non**

OCR - paramètres re-ranking

- Largeur de faisceau : 1, 2, 5, **10**, 20
- Coefficient de réduction pour mots inconnus : **0,75**



- Prise en compte de la fréquence ? **Non**

OCR - Préparation des corpus d'entraînement

Corpus Occitan

Annotateurs	Marianne Vergez-Couret
Nombre de livres numérisés	10
Années d'édition	1960-2000
Lieu d'édition	France
Nombre de pages	80
Nombre de mots	20 400
Nombre de lettres	85 500

Correction des erreurs humaines d'annotation après une première analyse automatique du corpus

OCR - Préparation des corpus d'entraînement

Corpus Occitan

Annotateurs	Marianne Vergez-Couret
Nombre de livres numérisés	10
Années d'édition	1960-2000
Lieu d'édition	France
Nombre de pages	80
Nombre de mots	20 400
Nombre de lettres	85 500

Correction des erreurs humaines d'annotation après une première analyse automatique du corpus

OCR - Ressources lexicales pour l'occitan

De quelles ressources dispose-t-on ?

- Textes BaTelÒc → lexiques de fléchies
- Dictionnaires sous format papier numérisés → lexiques de lemmes avec des informations grammaticales
  - Nombre d'entrée : 54 500
  - Génération des formes fléchies : 84 400

OCR - Ressources lexicales pour l'occitan

De quelles ressources dispose-t-on ?

- Textes BaTelÒc $\rightarrow$ lexiques de fléchies
- Dictionnaires sous format papier numérisés $\rightarrow$ lexiques de lemmes avec des informations grammaticales
  - Nombre d'entrée : 54 500
  - Génération des formes fléchies : 84 400

OCR - Ressources lexicales pour l'occitan

De quelles ressources dispose-t-on ?

- Textes BaTelÒc → lexiques de fléchies
- Dictionnaires sous format papier numérisés → lexiques de lemmes avec des informations grammaticales
  - Nombre d'entrée : 54 500
  - Génération des formes fléchies : 84 400

OCR - Ressources lexicales pour l'occitan

De quelles ressources dispose-t-on ?

- Textes BaTelÒc → lexiques de fléchies
- Dictionnaires sous format papier numérisés → lexiques de lemmes avec des informations grammaticales
  - Nombre d'entrée : 54 500
  - Génération des formes fléchies : 84 400

OCR - Ressources lexicales pour l'occitan

De quelles ressources dispose-t-on ?

- Textes BaTelÒc $\rightarrow$ lexiques de fléchies
- Dictionnaires sous format papier numérisés $\rightarrow$ lexiques de lemmes avec des informations grammaticales
  - Nombre d'entrée : 54 500
  - Génération des formes fléchies : 84 400

OCR - Fusion des lexiques

- Lex_Global (150 700)
  - Lex_Lengadocian (135 300)
    - Lex_Rouquette (17 100)
    - Lex_Laux (84 800)
    - Lex_Molin (9 600)
  - Lex_Gascon (28 900)
    - Lex_Blader (5 300)

OCR - Résultats

	Ensemble du corpus		Corpus Lengadocian		Corpus Gascon	
	Mots	Lettres	Mots	Lettres	Mots	Lettres
Sans Lexique	91,54	97,53	92,08	97,64	90,99	97,41
Lex_Gascon	92,72	97,81	93,07	97,85	92,36	97,78
Lex_Lengadocian	92,83	97,86	94,10	98,15	91,53	97,56
Lex_Global	**93,13**	**97,93**	94,08	98,13	92,16	97,71

Pour le corpus global :

- Apport systématique des lexiques
- Meilleure stratégie : Lex_Global
- Gain de 19% (mots) et 16% (lettres)

OCR - Résultats

	Ensemble du corpus		Corpus Lengadocian		Corpus Gascon	
	Mots	Lettres	Mots	Lettres	Mots	Lettres
Sans Lexique	91,54	97,53	92,08	97,64	90,99	97,41
Lex_Gascon	92,72	97,81	93,07	97,85	92,36	97,78
Lex_Lengadocian	92,83	97,86	94,10	98,15	91,53	97,56
Lex_Global	**93,13**	**97,93**	94,08	98,13	92,16	97,71

Pour le corpus global :

- Apport systématique des lexiques
- Meilleure stratégie : Lex_Global
- Gain de 19% (mots) et 16% (lettres)

OCR - Résultats

	Ensemble du corpus		Corpus Lengadocian		Corpus Gascon	
	Mots	Lettres	Mots	Lettres	Mots	Lettres
Sans Lexique	91,54	97,53	92,08	97,64	90,99	97,41
Lex_Gascon	92,72	97,81	93,07	97,85	92,36	97,78
Lex_Lengadocian	92,83	97,86	**94,10**	**98,15**	91,53	97,56
Lex_Global	93,13	97,93	94,08	98,13	92,16	97,71

Pour le sous-corpus lengadocian :

- Meilleure stratégie : Lexique du lengadocian
- Gain de 25,5% (mots) et 21,6% (lettres)
- Lexique du gascon = gain de 6%

OCR - Résultats

	Ensemble du corpus		Corpus Lengadocian		Corpus Gascon	
	Mots	Lettres	Mots	Lettres	Mots	Lettres
Sans Lexique	91,54	97,53	92,08	97,64	90,99	97,41
Lex_Gascon	92,72	97,81	93,07	97,85	92,36	97,78
Lex_Lengadocian	92,83	97,86	**94,10**	**98,15**	91,53	97,56
Lex_Global	93,13	97,93	94,08	98,13	92,16	97,71

Pour le sous-corpus lengadocian :

- Meilleure stratégie : Lexique du lengadocian
- Gain de 25,5% (mots) et 21,6% (lettres)
- Lexique du gascon = gain de 6%

OCR - Résultats

	Ensemble du corpus		Corpus Lengadocian		Corpus Gascon	
	Mots	Lettres	Mots	Lettres	Mots	Lettres
Sans Lexique	91,54	97,53	92,08	97,64	90,99	97,41
Lex_Gascon	92,72	97,81	93,07	97,85	**92,36**	**97,78**
Lex_Lengadocian	92,83	97,86	94,10	98,15	91,53	97,56
Lex_Global	93,13	97,93	94,08	98,13	92,16	97,71

Pour le sous-corpus gascon :

- Meilleure stratégie : Lexique du gascon
- Gain de 15,2% (mots) et 14,2% (lettres)
- Lexique du lengadocian = gain de 12%

OCR - Résultats

	Ensemble du corpus		Corpus Lengadocian		Corpus Gascon	
	Mots	Lettres	Mots	Lettres	Mots	Lettres
Sans Lexique	91,54	97,53	92,08	97,64	90,99	97,41
Lex_Gascon	92,72	97,81	93,07	97,85	**92,36**	**97,78**
Lex_Lengadocian	92,83	97,86	94,10	98,15	91,53	97,56
Lex_Global	93,13	97,93	94,08	98,13	92,16	97,71

Pour le sous-corpus gascon :

- Meilleure stratégie : Lexique du gascon
- Gain de 15,2% (mots) et 14,2% (lettres)
- Lexique du lengadocian = gain de 12%

OCR - Conclusion et perspectives

- Rôle des lexiques dans la tâche d'océrisation
  - Apport des lexiques par dialectes (occitan)
- Effort lexical
  - Difficultés intrinsèques à reconnaître les mots inconnus
  - Analyse et typologie des mots inconnus
  - Pour viser l'effort lexical le plus utile
  - Quelle est notre marge d'amélioration ?
- Pour le reste, effort d'ingénierie (segmentation, traits spécialisés)
- Apprentissage d'un modèle par police ? (italique, . . .)

OCR - Conclusion et perspectives

- Rôle des lexiques dans la tâche d'océrisation
  - Apport des lexiques par dialectes (occitan)
- Effort lexical
  - Difficultés intrinsèques à reconnaître les mots inconnus
  - Analyse et typologie des mots inconnus
  - Pour viser l'effort lexical le plus utile
  - Quelle est notre marge d'amélioration ?
- Pour le reste, effort d'ingénierie (segmentation, traits spécialisés)
- Apprentissage d'un modèle par police ? (italique, . . .)

OCR - Conclusion et perspectives

- Rôle des lexiques dans la tâche d'océrisation
  - Apport des lexiques par dialectes (occitan)
- Effort lexical
  - Difficultés intrinsèques à reconnaître les mots inconnus
  - Analyse et typologie des mots inconnus
  - Pour viser l'effort lexical le plus utile
  - Quelle est notre marge d'amélioration ?
- Pour le reste, effort d'ingénierie (segmentation, traits spécialisés)
- Apprentissage d'un modèle par police ? (italique, . . .)

Pos-tagging : Lengadocian and Gascon dialects

- Examples of lexical variations : filh/hilh ; luna/lua ; cabra/craba
- Examples of syntactic variations :
  - Enonciative particles
    - Example : "I'm buying bread and apples".
    - Gascon : *"**Que** crompi pans e pomas."*
    - Lengadocian : *"Compri de pans e de pomas."*
  - Indefinite and partitive articles
    - Example : "I want some water."
    - Gascon : *"Que vòli aiga."*
    - Lengadocian : *"Vòli **d'**aiga."*
  - Double/triple negation mandatory
    - Example : "He can't hear anything."
    - Gascon : *"**N'**enten pas arren."*
    - Lengadocian : *"Enten pas ren."*
- Additional intra-dialectal and spelling variations

Pos-tagging : Lengadocian and Gascon dialects

- Examples of lexical variations : filh/hilh ; luna/lua ; cabra/craba
- Examples of syntactic variations :
  - Enonciative particles
    - Example : "I'm buying bread and apples".
    - Gascon : "**Que** *crompi pans e pomas.*"
    - Lengadocian : *"Compri de pans e de pomas."*
  - Indefinite and partitive articles
    - Example : "I want some water."
    - Gascon : *"Que vòli aiga."*
    - Lengadocian : *"Vòli* **d'***aiga."*
  - Double/triple negation mandatory
    - Example : "He can't hear anything."
    - Gascon : *"***N'***enten pas arren."*
    - Lengadocian : *"Enten pas ren."*
- Additional intra-dialectal and spelling variations

Pos-tagging : Lengadocian and Gascon dialects

- Examples of lexical variations : filh/hilh ; luna/lua ; cabra/craba
- Examples of syntactic variations :
  - Enonciative particles
    - Example : "I'm buying bread and apples".
    - Gascon : *"***Que** *crompi pans e pomas."*
    - Lengadocian : *"Compri de pans e de pomas."*
  - Indefinite and partitive articles
    - Example : "I want some water."
    - Gascon : *"Que vòli aiga."*
    - Lengadocian : *"Vòli* **d'***aiga."*
  - Double/triple negation mandatory
    - Example : "He can't hear anything."
    - Gascon : *"***N'***enten pas arren."*
    - Lengadocian : *"Enten pas ren."*
- Additional intra-dialectal and spelling variations

Pos-tagging : Lengadocian and Gascon dialects

- Examples of lexical variations : filh/hilh ; luna/lua ; cabra/craba
- Examples of syntactic variations :
  - Enonciative particles
    - Example : "I'm buying bread and apples".
    - Gascon : "**Que** *crompi pans e pomas.*"
    - Lengadocian : *"Compri de pans e de pomas."*
  - Indefinite and partitive articles
    - Example : "I want some water."
    - Gascon : *"Que vòli aiga."*
    - Lengadocian : *"Vòli* **d'***aiga."*
  - Double/triple negation mandatory
    - Example : "He can't hear anything."
    - Gascon : "**N'***enten pas arren."*
    - Lengadocian : *"Enten pas ren."*
- Additional intra-dialectal and spelling variations

Pos-tagging : Lengadocian and Gascon dialects

- Examples of lexical variations : filh/hilh ; luna/lua ; cabra/craba
- Examples of syntactic variations :
  - Enonciative particles
    - Example : "I'm buying bread and apples".
    - Gascon : "**Que** *crompi pans e pomas.*"
    - Lengadocian : *"Compri de pans e de pomas."*
  - Indefinite and partitive articles
    - Example : "I want some water."
    - Gascon : *"Que vòli aiga."*
    - Lengadocian : *"Vòli* **d'***aiga."*
  - Double/triple negation mandatory
    - Example : "He can't hear anything."
    - Gascon : "**N'***enten pas arren."*
    - Lengadocian : *"Enten pas ren."*
- Additional intra-dialectal and spelling variations

Pos-tagging : Software

- Talismane (Urieli, 2013)
  - Supervised machine learning approach
  - Linear start-to-end pos-tagging
  - Open source
  - http ://redac.univ-tlse2.fr/talismane.html
  - $\approx$ 97% accuracy on English and French
- Lexicon usage :
  - As features, to help the statistical model
  - As rules, to override the statistical model
- Machine learning :
  - Linear SVM
  - Parameters : $\epsilon = 0.1$, $C = 0.5$

Pos-tagging : Software

- Talismane (Urieli, 2013)
  - Supervised machine learning approach
  - Linear start-to-end pos-tagging
  - Open source
  - http ://redac.univ-tlse2.fr/talismane.html
  - $\approx$ 97% accuracy on English and French
- Lexicon usage :
  - As features, to help the statistical model
  - As rules, to override the statistical model
- Machine learning :
  - Linear SVM
  - Parameters : $\epsilon = 0.1$, $C = 0.5$

Pos-tagging : Software

- Talismane (Urieli, 2013)
  - Supervised machine learning approach
  - Linear start-to-end pos-tagging
  - Open source
  - http ://redac.univ-tlse2.fr/talismane.html
  - $\approx$ 97% accuracy on English and French
- Lexicon usage :
  - As features, to help the statistical model
  - As rules, to override the statistical model
- Machine learning :
  - Linear SVM
  - Parameters : $\epsilon = 0.1$, $C = 0.5$

Features

- Same base features as for English and French :
  - W : the word form
  - P : the pos-tag (assigned or in lexicon)
  - L : the lemma (assigned or in lexicon)
  - U : if the token is unknown in the lexicon
  - $Pref_n$ / Sfx_n : the first/last n letters
  - *1st* / *Last* : if the token is 1st/last in the sentence
  - 2- and 3-grams built from tokens at positions -2, -1, 0, +1, +2

Features

- Same base features as for English and French :
  - *W* : the word form
  - *P* : the pos-tag (assigned or in lexicon)
  - *L* : the lemma (assigned or in lexicon)
  - *U* : if the token is unknown in the lexicon
  - $Pref_n$ / Sfx_n : the first/last n letters
  - *1st* / *Last* : if the token is 1st/last in the sentence
  - 2- and 3-grams built from tokens at positions -2, -1, 0, +1, +2

Features

- Same base features as for English and French :
  - W : the word form
  - P : the pos-tag (assigned or in lexicon)
  - L : the lemma (assigned or in lexicon)
  - U : if the token is unknown in the lexicon
  - $Pref_n$ / Sfx_n : the first/last n letters
  - *1st* / *Last* : if the token is 1st/last in the sentence
  - 2- and 3-grams built from tokens at positions -2, -1, 0, +1, +2

Features

- Same base features as for English and French :
  - W : the word form
  - P : the pos-tag (assigned or in lexicon)
  - L : the lemma (assigned or in lexicon)
  - U : if the token is unknown in the lexicon
  - $Pref_n$ / Sfx_n : the first/last n letters
  - *1st* / *Last* : if the token is 1st/last in the sentence
  - 2- and 3-grams built from tokens at positions -2, -1, 0, +1, +2

Features

- Same base features as for English and French :
  - *W* : the word form
  - *P* : the pos-tag (assigned or in lexicon)
  - *L* : the lemma (assigned or in lexicon)
  - *U* : if the token is unknown in the lexicon
  - $Pref_n$ / Sfx_n : the first/last n letters
  - *1st* / *Last* : if the token is 1st/last in the sentence
  - 2- and 3-grams built from tokens at positions -2, -1, 0, +1, +2

Features

- Same base features as for English and French :
  - *W* : the word form
  - *P* : the pos-tag (assigned or in lexicon)
  - *L* : the lemma (assigned or in lexicon)
  - *U* : if the token is unknown in the lexicon
  - $Pref_n$ / Sfx_n : the first/last n letters
  - *1st* / *Last* : if the token is 1st/last in the sentence
  - 2- and 3-grams built from tokens at positions -2, -1, 0, +1, +2

Features

- Same base features as for English and French :
  - *W* : the word form
  - *P* : the pos-tag (assigned or in lexicon)
  - *L* : the lemma (assigned or in lexicon)
  - *U* : if the token is unknown in the lexicon
  - $Pref_n$ / Sfx_n : the first/last n letters
  - *1st* / *Last* : if the token is 1st/last in the sentence
  - 2- and 3-grams built from tokens at positions -2, -1, 0, +1, +2

Features

- Same base features as for English and French :
  - *W* : the word form
  - *P* : the pos-tag (assigned or in lexicon)
  - *L* : the lemma (assigned or in lexicon)
  - *U* : if the token is unknown in the lexicon
  - $Pref_n$ / Sfx_n : the first/last n letters
  - *1st* / *Last* : if the token is 1st/last in the sentence
  - 2- and 3-grams built from tokens at positions -2, -1, 0, +1, +2

Rules

- Rules from lexicon for :
  - Closed classes (non-productive functional categories)
  - Open classes (productive lexical categories)
- Three rules :
  1. Closed classes : only assign preposition, conjunction, etc. if the form is listed in the lexicon for this pos-tag = Don't invent new prepositions
  2. Open classes : don't assign common noun, adjective, etc. if the form is only listed with closed classes in lexicon = Don't assign common noun to *"lo"* ("the").
  3. Automatically assign `Card` and `Pct` respectively to numbers and punctuation.

Rules

- Rules from lexicon for :
  - Closed classes (non-productive functional categories)
  - Open classes (productive lexical categories)
- Three rules :
  1. Closed classes : only assign preposition, conjunction, etc. if the form is listed in the lexicon for this pos-tag = Don't invent new prepositions
  2. Open classes : don't assign common noun, adjective, etc. if the form is only listed with closed classes in lexicon = Don't assign common noun to *"lo"* ("the").
  3. Automatically assign `Card` and `Pct` respectively to numbers and punctuation.

Rules

- Rules from lexicon for :
  - Closed classes (non-productive functional categories)
  - Open classes (productive lexical categories)
- Three rules :
  1. Closed classes : only assign preposition, conjunction, etc. if the form is listed in the lexicon for this pos-tag = Don't invent new prepositions
  2. Open classes : don't assign common noun, adjective, etc. if the form is only listed with closed classes in lexicon = Don't assign common noun to *"lo"* ("the").
  3. Automatically assign `Card` and `Pct` respectively to numbers and punctuation.

Rules

- Rules from lexicon for :
  - Closed classes (non-productive functional categories)
  - Open classes (productive lexical categories)
- Three rules :
  1. Closed classes : only assign preposition, conjunction, etc. if the form is listed in the lexicon for this pos-tag = Don't invent new prepositions
  2. Open classes : don't assign common noun, adjective, etc. if the form is only listed with closed classes in lexicon = Don't assign common noun to *"lo"* ("the").
  3. Automatically assign `Card` and `Pct` respectively to numbers and punctuation.

Rules

- Rules from lexicon for :
  - Closed classes (non-productive functional categories)
  - Open classes (productive lexical categories)
- Three rules :
  1. Closed classes : only assign preposition, conjunction, etc. if the form is listed in the lexicon for this pos-tag = Don't invent new prepositions
  2. Open classes : don't assign common noun, adjective, etc. if the form is only listed with closed classes in lexicon = Don't assign common noun to *"lo"* ("the").
  3. Automatically assign `Card` and `Pct` respectively to numbers and punctuation.

Rules

- Rules from lexicon for :
  - Closed classes (non-productive functional categories)
  - Open classes (productive lexical categories)
- Three rules :
  1. Closed classes : only assign preposition, conjunction, etc. if the form is listed in the lexicon for this pos-tag = Don't invent new prepositions
  2. Open classes : don't assign common noun, adjective, etc. if the form is only listed with closed classes in lexicon = Don't assign common noun to *"lo"* ("the").
  3. Automatically assign `Card` and `Pct` respectively to numbers and punctuation.

Rules

- Rules from lexicon for :
  - Closed classes (non-productive functional categories)
  - Open classes (productive lexical categories)
- Three rules :
  1. Closed classes : only assign preposition, conjunction, etc. if the form is listed in the lexicon for this pos-tag = Don't invent new prepositions
  2. Open classes : don't assign common noun, adjective, etc. if the form is only listed with closed classes in lexicon = Don't assign common noun to *"lo"* ("the").
  3. Automatically assign `Card` and `Pct` respectively to numbers and punctuation.

Resources

- For Talismane to function properly, various resources are required :
  - A training corpus from which the statistical model is learned : **Lengadocian Training Corpus**
  - One or more evaluation corpora to evaluate performance : **2 Lengadocian (Rouergue and Lot), 1 Gascon**
  - Optionally a lexicon for wide-coverage features and rules : **Lengadocian Lexicon**
  - All rely on a tagset specifically designed for Occitan.

Resources

- For Talismane to function properly, various resources are required :
  - A training corpus from which the statistical model is learned : **Lengadocian Training Corpus**
  - One or more evaluation corpora to evaluate performance : **2 Lengadocian (Rouergue and Lot), 1 Gascon**
  - Optionally a lexicon for wide-coverage features and rules : **Lengadocian Lexicon**
  - All rely on a tagset specifically designed for Occitan.

Resources

- For Talismane to function properly, various resources are required :
  - A training corpus from which the statistical model is learned : **Lengadocian Training Corpus**
  - One or more evaluation corpora to evaluate performance : **2 Lengadocian (Rouergue and Lot), 1 Gascon**
  - Optionally a lexicon for wide-coverage features and rules : **Lengadocian Lexicon**
  - All rely on a tagset specifically designed for Occitan.

Resources

- For Talismane to function properly, various resources are required :
  - A training corpus from which the statistical model is learned : **Lengadocian Training Corpus**
  - One or more evaluation corpora to evaluate performance : **2 Lengadocian (Rouergue and Lot), 1 Gascon**
  - Optionally a lexicon for wide-coverage features and rules : **Lengadocian Lexicon**
  - All rely on a tagset specifically designed for Occitan.

Resources

- For Talismane to function properly, various resources are required :
  - A training corpus from which the statistical model is learned : **Lengadocian Training Corpus**
  - One or more evaluation corpora to evaluate performance : **2 Lengadocian (Rouergue and Lot), 1 Gascon**
  - Optionally a lexicon for wide-coverage features and rules : **Lengadocian Lexicon**
  - All rely on a tagset specifically designed for Occitan.

Tagset

Tag	Description	Lexicon size
A	Adjective	29,638
Adv	Adverb	751
Cc	Coordinating conjunction	8
Cs	Subordinating conjunction	150
Det	Article	127
Card	Cardinal number	42
Cli	Clitic	72
CliRef	Reflexive pronoun	17
Inj	Interjection	7
Nc	Common noun	25,817
Np	Proper noun	4,603
Pct	Punctuation	15
Pe	Enunciative particle (Gascon only)	0
Pp	Present participle	4,530
Pr	Preposition	521
Prel	Relative pronoun	37
Pro	Pronoun	81
Ps	Past participle	17,963
PrepDet	Amalgamated preposition and article	499
Vc	Conjugated verb	135,731
Vi	Infinitive verb	4,643
Z	Consonant for phonetic liaison	3
Total		**225,386**

Training Corpus

- Lengadocian Dialect - Rouergue Varieties from *E la barta floriguèt* by Enric Molin
- 2500 tokens (lemma + pos-tags)

Index	Token	Lemma	Pos-tag	Morphology
1	Li	li	Cli	P3-m-sg
2	semblava	semblar	Vc	Imi-P3-sg
3	que	que	Cs	
4	sos	son	D	Poss-P3-m-pl
5	pès	pè	Nc	m-pl
6	tocavan	tocar	Vc	Imi-P3-pl
7	pas	pas	Adv	
8	tèrra	tèrra	Nc	f-sg
9	.	.	Pct	

Training Corpus

- Lengadocian Dialect - Rouergue Varieties from *E la barta floriguèt* by Enric Molin
- 2500 tokens (lemma + pos-tags)

Index	Token	Lemma	Pos-tag	Morphology
1	Li	li	Cli	P3-m-sg
2	semblava	semblar	Vc	Imi-P3-sg
3	que	que	Cs	
4	sos	son	D	Poss-P3-m-pl
5	pès	pè	Nc	m-pl
6	tocavan	tocar	Vc	Imi-P3-pl
7	pas	pas	Adv	
8	tèrra	tèrra	Nc	f-sg
9	.	.	Pct	

Training Corpus

- Lengadocian Dialect - Rouergue Varieties from *E la barta floriguèt* by Enric Molin
- 2500 tokens (lemma + pos-tags)

Index	Token	Lemma	Pos-tag	Morphology
1	Li	li	Cli	P3-m-sg
2	semblava	semblar	Vc	Imi-P3-sg
3	que	que	Cs	
4	sos	son	D	Poss-P3-m-pl
5	pès	pè	Nc	m-pl
6	tocavan	tocar	Vc	Imi-P3-pl
7	pas	pas	Adv	
8	tèrra	tèrra	Nc	f-sg
9	.	.	Pct	

Evaluation Corpora

- Lengadocian Dialect - Rouergue Varieties from *Los crocants de Roergue* by Ferran Delèris
  700 tokens (lemma + pos-tag)
- Lengadocian Dialect - Lot Varieties from *Dels camins bartassièrs* by Marceu Esquieu
  460 tokens (lemma + pos-tag)
- Gascon Dialect from *Hont Blanc* by Jan Loís Lavit
  460 tokens (lemma + pos-tag)

Evaluation Corpora

- Lengadocian Dialect - Rouergue Varieties from *Los crocants de Roergue* by Ferran Delèris
  700 tokens (lemma + pos-tag)
- Lengadocian Dialect - Lot Varieties from *Dels camins bartassièrs* by Marceu Esquieu
  460 tokens (lemma + pos-tag)
- Gascon Dialect from *Hont Blanc* by Jan Loís Lavit
  460 tokens (lemma + pos-tag)

Evaluation Corpora

- Lengadocian Dialect - Rouergue Varieties from *Los crocants de Roergue* by Ferran Delèris
  700 tokens (lemma + pos-tag)
- Lengadocian Dialect - Lot Varieties from *Dels camins bartassièrs* by Marceu Esquieu
  460 tokens (lemma + pos-tag)
- Gascon Dialect from *Hont Blanc* by Jan Loís Lavit
  460 tokens (lemma + pos-tag)

Corpus comparison

Corpus	Training	Rouergue	Lot	Gascon
Size	2501	701	467	469
Size (without punct.)	2078	591	388	399
% unknown in training		46.4%	49.0%	56.4%
% unknown in lexicon	0.1%	16.6%	19.9%	40.1%
Open class tokens	1111	324	201	203
% unknown in training		76.2%	82.6%	**87.7%**
% unknown in lexicon	0.2%	29.0%	37.3%	**59.1%**
Closed class tokens	967	267	187	196
% unknown in training		10.2%	12.8%	**24.0%**
% unknown in lexicon	0.0%	1.5%	1.1%	**20.4%**

TABLE: Training and evaluation corpora

Experiments

Questions for experiments :

- Which is the best strategy for each evaluation corpus ?
- Is it always useful to apply closed-class rules ?
- To what extent can a model built from a training corpus for a single dialectal variety be applied to other varieties and dialects ?
- To what extent can a lexicon for one dialect be applied to another dialect ?
- What methods can be used to improve analysis for a dialect different from the training/lexicon dialect ?
- Given limited resources, is it better to annotate a larger training corpus, or compile a larger lexicon ?

Experiments

Questions for experiments :

- Which is the best strategy for each evaluation corpus ?
- Is it always useful to apply closed-class rules ?
- To what extent can a model built from a training corpus for a single dialectal variety be applied to other varieties and dialects ?
- To what extent can a lexicon for one dialect be applied to another dialect ?
- What methods can be used to improve analysis for a dialect different from the training/lexicon dialect ?
- Given limited resources, is it better to annotate a larger training corpus, or compile a larger lexicon ?

Experiments

Questions for experiments :

- Which is the best strategy for each evaluation corpus ?
- Is it always useful to apply closed-class rules ?
- To what extent can a model built from a training corpus for a single dialectal variety be applied to other varieties and dialects ?
- To what extent can a lexicon for one dialect be applied to another dialect ?
- What methods can be used to improve analysis for a dialect different from the training/lexicon dialect ?
- Given limited resources, is it better to annotate a larger training corpus, or compile a larger lexicon ?

Experiments

Questions for experiments :

- Which is the best strategy for each evaluation corpus ?
- Is it always useful to apply closed-class rules ?
- To what extent can a model built from a training corpus for a single dialectal variety be applied to other varieties and dialects ?
- To what extent can a lexicon for one dialect be applied to another dialect ?
- What methods can be used to improve analysis for a dialect different from the training/lexicon dialect ?
- Given limited resources, is it better to annotate a larger training corpus, or compile a larger lexicon ?

Experiments

Questions for experiments :

- Which is the best strategy for each evaluation corpus ?
- Is it always useful to apply closed-class rules ?
- To what extent can a model built from a training corpus for a single dialectal variety be applied to other varieties and dialects ?
- To what extent can a lexicon for one dialect be applied to another dialect ?
- What methods can be used to improve analysis for a dialect different from the training/lexicon dialect ?
- Given limited resources, is it better to annotate a larger training corpus, or compile a larger lexicon ?

Experiments

Questions for experiments :

- Which is the best strategy for each evaluation corpus ?
- Is it always useful to apply closed-class rules ?
- To what extent can a model built from a training corpus for a single dialectal variety be applied to other varieties and dialects ?
- To what extent can a lexicon for one dialect be applied to another dialect ?
- What methods can be used to improve analysis for a dialect different from the training/lexicon dialect ?
- Given limited resources, is it better to annotate a larger training corpus, or compile a larger lexicon ?

Experiments

Questions for experiments :

- Which is the best strategy for each evaluation corpus ?
- Is it always useful to apply closed-class rules ?
- To what extent can a model built from a training corpus for a single dialectal variety be applied to other varieties and dialects ?
- To what extent can a lexicon for one dialect be applied to another dialect ?
- What methods can be used to improve analysis for a dialect different from the training/lexicon dialect ?
- Given limited resources, is it better to annotate a larger training corpus, or compile a larger lexicon ?

Overall Results

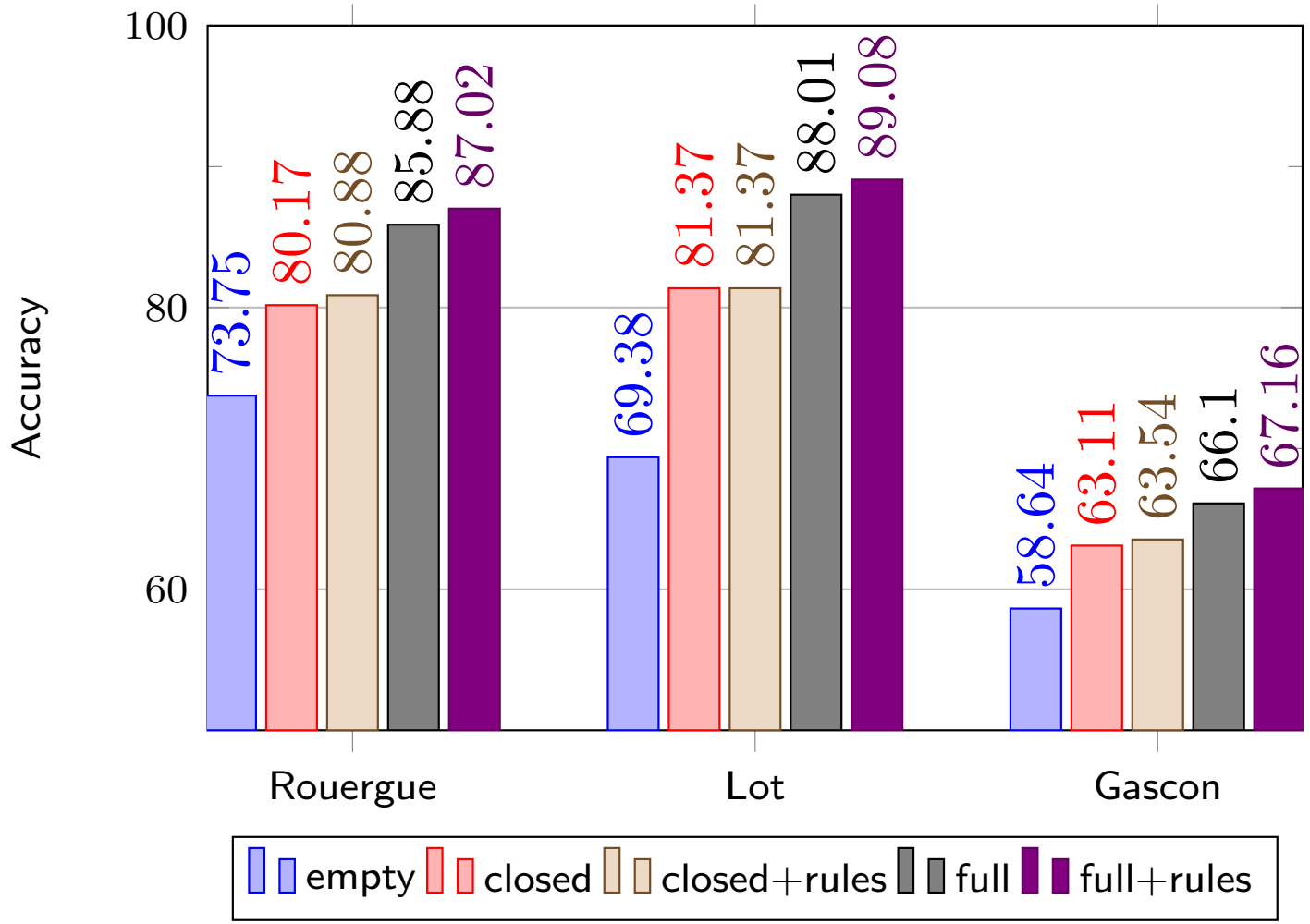


FIGURE: Pos-tagging lexicon/rules comparison : accuracy by corpus

Closed Class Rules

- When adding closed class rules :
  - Rouergue : 85.88% to 87.02%
  - Lot : 88.01% to 89.08%
  - Gascon : 66.10% to 67.16%
- Always helpful, even for a corpus (Gascon) with 20% unknown closed class tokens

Closed Class Rules

- When adding closed class rules :
  - Rouergue : 85.88% to 87.02%
  - Lot : 88.01% to 89.08%
  - Gascon : 66.10% to 67.16%
- Always helpful, even for a corpus (Gascon) with 20% unknown closed class tokens

Closed Class Rules

- When adding closed class rules :
  - Rouergue : 85.88% to 87.02%
  - Lot : 88.01% to 89.08%
  - Gascon : 66.10% to 67.16%
- Always helpful, even for a corpus (Gascon) with 20% unknown closed class tokens

Lexicons

- Gain : no lexicon → closed-class lexicon
  - Rouergue : 7.13%
  - Lot : 11.99%
  - Gascon : 4.90%
- Gain : closed-class lexicon → full lexicon
  - Rouergue : 6.14%
  - Lot : 7.71%
  - Gascon : 3.62%
- Mean gain for unknown words (mostly through n-grams) : half lexicon → full lexicon
  - Rouergue : 8.54%
  - Lot : 17.96%

Lexicons

- Gain : no lexicon → closed-class lexicon
  - Rouergue : 7.13%
  - Lot : 11.99%
  - Gascon : 4.90%
- Gain : closed-class lexicon → full lexicon
  - Rouergue : 6.14%
  - Lot : 7.71%
  - Gascon : 3.62%
- Mean gain for unknown words (mostly through n-grams) : half lexicon → full lexicon
  - Rouergue : 8.54%
  - Lot : 17.96%

Lexicons

- Gain : no lexicon → closed-class lexicon
  - Rouergue : 7.13%
  - Lot : 11.99%
  - Gascon : 4.90%
- Gain : closed-class lexicon → full lexicon
  - Rouergue : 6.14%
  - Lot : 7.71%
  - Gascon : 3.62%
- Mean gain for unknown words (mostly through n-grams) : half lexicon → full lexicon
  - Rouergue : 8.54%
  - Lot : 17.96%

Other Dialects

- Gascon : new part-of-speech, the enunciative particle (Pe)
- Most common for *"que"*, only possibility for *"be"*
- New rule :
  - Always annotate *"be"* as Pe
  - Annotate *"que"* as Pe at start-of-sentence, after conjunction and after comma
  - Result : 17 true positives, 1 false positive, 13 false negatives
  - F-score = 70.83%, Total accuracy from 67.16% to 69.72%
- Next steps for dialect
  - More rules, full closed-class lexicon for Gascon, training corpus for Gascon
  - Better to use lexicon per dialect or full lexicon ?
  - Better to use training corpus per dialect or full training corpus ?

Other Dialects

- Gascon : new part-of-speech, the enunciative particle (Pe)
- Most common for *"que"*, only possibility for *"be"*
- New rule :
  - Always annotate *"be"* as Pe
  - Annotate *"que"* as Pe at start-of-sentence, after conjunction and after comma
  - Result : 17 true positives, 1 false positive, 13 false negatives
  - F-score = 70.83%, Total accuracy from 67.16% to 69.72%
- Next steps for dialect
  - More rules, full closed-class lexicon for Gascon, training corpus for Gascon
  - Better to use lexicon per dialect or full lexicon ?
  - Better to use training corpus per dialect or full training corpus ?

Other Dialects

- Gascon : new part-of-speech, the enunciative particle (Pe)
- Most common for *"que"*, only possibility for *"be"*
- New rule :
  - Always annotate *"be"* as Pe
  - Annotate *"que"* as Pe at start-of-sentence, after conjunction and after comma
  - Result : 17 true positives, 1 false positive, 13 false negatives
  - F-score = 70.83%, Total accuracy from 67.16% to 69.72%
- Next steps for dialect
  - More rules, full closed-class lexicon for Gascon, training corpus for Gascon
  - Better to use lexicon per dialect or full lexicon ?
  - Better to use training corpus per dialect or full training corpus ?

Other Dialects

- Gascon : new part-of-speech, the enunciative particle (Pe)
- Most common for *"que"*, only possibility for *"be"*
- New rule :
  - Always annotate *"be"* as Pe
  - Annotate *"que"* as Pe at start-of-sentence, after conjunction and after comma
  - Result : 17 true positives, 1 false positive, 13 false negatives
  - F-score = 70.83%, Total accuracy from 67.16% to 69.72%
- Next steps for dialect
  - More rules, full closed-class lexicon for Gascon, training corpus for Gascon
  - Better to use lexicon per dialect or full lexicon ?
  - Better to use training corpus per dialect or full training corpus ?

Other Dialects

- Gascon : new part-of-speech, the enunciative particle (Pe)
- Most common for *"que"*, only possibility for *"be"*
- New rule :
  - Always annotate *"be"* as Pe
  - Annotate *"que"* as Pe at start-of-sentence, after conjunction and after comma
  - Result : 17 true positives, 1 false positive, 13 false negatives
  - F-score = 70.83%, Total accuracy from 67.16% to 69.72%
- Next steps for dialect
  - More rules, full closed-class lexicon for Gascon, training corpus for Gascon
  - Better to use lexicon per dialect or full lexicon ?
  - Better to use training corpus per dialect or full training corpus ?

Other Dialects

- Gascon : new part-of-speech, the enunciative particle (Pe)
- Most common for *"que"*, only possibility for *"be"*
- New rule :
  - Always annotate *"be"* as Pe
  - Annotate *"que"* as Pe at start-of-sentence, after conjunction and after comma
  - Result : 17 true positives, 1 false positive, 13 false negatives
  - F-score = 70.83%, Total accuracy from 67.16% to 69.72%
- Next steps for dialect
  - More rules, full closed-class lexicon for Gascon, training corpus for Gascon
  - Better to use lexicon per dialect or full lexicon ?
  - Better to use training corpus per dialect or full training corpus ?

Other Dialects

- Gascon : new part-of-speech, the enunciative particle (Pe)
- Most common for *"que"*, only possibility for *"be"*
- New rule :
  - Always annotate *"be"* as Pe
  - Annotate *"que"* as Pe at start-of-sentence, after conjunction and after comma
  - Result : 17 true positives, 1 false positive, 13 false negatives
  - F-score = 70.83%, Total accuracy from 67.16% to 69.72%
- Next steps for dialect
  - More rules, full closed-class lexicon for Gascon, training corpus for Gascon
  - Better to use lexicon per dialect or full lexicon ?
  - Better to use training corpus per dialect or full training corpus ?

Other Dialects

- Gascon : new part-of-speech, the enunciative particle (Pe)
- Most common for *"que"*, only possibility for *"be"*
- New rule :
  - Always annotate *"be"* as Pe
  - Annotate *"que"* as Pe at start-of-sentence, after conjunction and after comma
  - Result : 17 true positives, 1 false positive, 13 false negatives
  - F-score = 70.83%, Total accuracy from 67.16% to 69.72%
- Next steps for dialect
  - More rules, full closed-class lexicon for Gascon, training corpus for Gascon
  - Better to use lexicon per dialect or full lexicon ?
  - Better to use training corpus per dialect or full training corpus ?

Other Dialects

- Gascon : new part-of-speech, the enunciative particle (Pe)
- Most common for *"que"*, only possibility for *"be"*
- New rule :
  - Always annotate *"be"* as Pe
  - Annotate *"que"* as Pe at start-of-sentence, after conjunction and after comma
  - Result : 17 true positives, 1 false positive, 13 false negatives
  - F-score = 70.83%, Total accuracy from 67.16% to 69.72%
- Next steps for dialect
  - More rules, full closed-class lexicon for Gascon, training corpus for Gascon
  - Better to use lexicon per dialect or full lexicon ?
  - Better to use training corpus per dialect or full training corpus ?

Other Dialects

- Gascon : new part-of-speech, the enunciative particle (Pe)
- Most common for *"que"*, only possibility for *"be"*
- New rule :
  - Always annotate *"be"* as Pe
  - Annotate *"que"* as Pe at start-of-sentence, after conjunction and after comma
  - Result : 17 true positives, 1 false positive, 13 false negatives
  - F-score = 70.83%, Total accuracy from 67.16% to 69.72%
- Next steps for dialect
  - More rules, full closed-class lexicon for Gascon, training corpus for Gascon
  - Better to use lexicon per dialect or full lexicon ?
  - Better to use training corpus per dialect or full training corpus ?

Training Corpus vs. Lexicon

- Given limited time, should we annotate more training corpus or build a larger lexicon ?
- Experiment : Create 2 lexicon halves, 2 training corpus halves.
  - Mean gain when doubling training corpus from 1250 to 2500 tokens : 1.46%
  - Mean gain when doubling the lexicon from 110K to 220K entries : 4.16%
- But : can always annotate more data, finding more lexical items more difficult

Training Corpus vs. Lexicon

- Given limited time, should we annotate more training corpus or build a larger lexicon ?
- Experiment : Create 2 lexicon halves, 2 training corpus halves.
  - Mean gain when doubling training corpus from 1250 to 2500 tokens : 1.46%
  - Mean gain when doubling the lexicon from 110K to 220K entries : 4.16%
- But : can always annotate more data, finding more lexical items more difficult

Training Corpus vs. Lexicon

- Given limited time, should we annotate more training corpus or build a larger lexicon ?
- Experiment : Create 2 lexicon halves, 2 training corpus halves.
  - Mean gain when doubling training corpus from 1250 to 2500 tokens : 1.46%
  - Mean gain when doubling the lexicon from 110K to 220K entries : 4.16%
- But : can always annotate more data, finding more lexical items more difficult

Training Corpus vs. Lexicon

- Given limited time, should we annotate more training corpus or build a larger lexicon ?
- Experiment : Create 2 lexicon halves, 2 training corpus halves.
  - Mean gain when doubling training corpus from 1250 to 2500 tokens : 1.46%
  - Mean gain when doubling the lexicon from 110K to 220K entries : 4.16%
- But : can always annotate more data, finding more lexical items more difficult

Training Corpus vs. Lexicon

- Given limited time, should we annotate more training corpus or build a larger lexicon ?
- Experiment : Create 2 lexicon halves, 2 training corpus halves.
  - Mean gain when doubling training corpus from 1250 to 2500 tokens : 1.46%
  - Mean gain when doubling the lexicon from 110K to 220K entries : 4.16%
- But : can always annotate more data, finding more lexical items more difficult

Pos-tagging : Conclusion and perspectives

- Reasonable results ($> 89\%$) with very little annotated material (2500 tokens), **if** wide-coverage lexicon is available
- It is better to construct a larger lexicon than to annotate more training material
- Functioning pos-tagger + annotation guide
- Cross-dialect pos-tagging (in our case, Gascon)
  - Rules (e.g. for enunciative particle)
  - Complete closed-class lexicon
  - Open-class lexicon + training corpus
  - But : separate by dialect or not ?
- Semi-supervised cross-language methods (Catalan) : more gains ?

Pos-tagging : Conclusion and perspectives

- Reasonable results ($> 89\%$) with very little annotated material (2500 tokens), **if** wide-coverage lexicon is available
- It is better to construct a larger lexicon than to annotate more training material
- Functioning pos-tagger + annotation guide
- Cross-dialect pos-tagging (in our case, Gascon)
  - Rules (e.g. for enunciative particle)
  - Complete closed-class lexicon
  - Open-class lexicon + training corpus
  - But : separate by dialect or not ?
- Semi-supervised cross-language methods (Catalan) : more gains ?

Pos-tagging : Conclusion and perspectives

- Reasonable results ($> 89\%$) with very little annotated material (2500 tokens), **if** wide-coverage lexicon is available
- It is better to construct a larger lexicon than to annotate more training material
- Functioning pos-tagger + annotation guide
- Cross-dialect pos-tagging (in our case, Gascon)
  - Rules (e.g. for enunciative particle)
  - Complete closed-class lexicon
  - Open-class lexicon + training corpus
  - But : separate by dialect or not ?
- Semi-supervised cross-language methods (Catalan) : more gains ?

Pos-tagging : Conclusion and perspectives

- Reasonable results ($> 89\%$) with very little annotated material (2500 tokens), **if** wide-coverage lexicon is available
- It is better to construct a larger lexicon than to annotate more training material
- Functioning pos-tagger + annotation guide
- Cross-dialect pos-tagging (in our case, Gascon)
  - Rules (e.g. for enunciative particle)
  - Complete closed-class lexicon
  - Open-class lexicon + training corpus
  - But : separate by dialect or not ?
- Semi-supervised cross-language methods (Catalan) : more gains ?

Pos-tagging : Conclusion and perspectives

- Reasonable results ($> 89\%$) with very little annotated material (2500 tokens), **if** wide-coverage lexicon is available
- It is better to construct a larger lexicon than to annotate more training material
- Functioning pos-tagger + annotation guide
- Cross-dialect pos-tagging (in our case, Gascon)
  - Rules (e.g. for enunciative particle)
  - Complete closed-class lexicon
  - Open-class lexicon + training corpus
  - But : separate by dialect or not ?
- Semi-supervised cross-language methods (Catalan) : more gains ?

❺ Conclusions et perspectives

- Méthodes et résultats encourageants pour démarrer RESTAURE
- Perfectionner les outils (OCR)
- Aller plus loin en exploitant les ressources des langues proches
- Aller plus loin dans la gestion des dialectes et cie (pour créer des systèmes robustes)
- Trouver des méthodes pour adapter les ressources disponibles aux différentes variantes