DE BATELÒC AU TALÒC

Ressources et outils pour le traitement de langues dites « peu dotées »

Marianne Vergez-Couret
Séminaire CLLE-ERSS
25/02/2015

Remerciements

- Aux collègues des projets BaTelÒc et RESTAURE : Myriam Bras, Nabil Hathout, Mai Ho-Dac, Franck Sajous, Jean Sibille, Assaf Urieli, Delphine Bernhard
- Aux collègues du Congrès Permanent de la lenga occitana et du CIRDOC
- Aux stagiaires
  - Estel Llansana, Sébastien Gonzales (Centre de formation professionnelle occitan)
  - Laurent Gilard, Estelle Pompon (Stage Master Culture et Patrimoine en Pays d'Oc)
  - Aurélie Abadie (L3 Occitan)
  - Lucie Berge et Eunbee Kang (L3 Sciences du langage)

Projets

- 2012-2014

  BaTelÒc (Région Midi-Pyrénées + Univ Tlse 2)

  Base de **Te**xtes pour la **l**angue d'**Òc**

  - Passage de la base expérimentale à la base opérationnelle
  - Passage des textes nus à des textes enrichis d'annotations morphosyntaxiques

- 2014-2015

  Projet ANR RESTAURE

  RESsources et outils pour le Traitement AUtomatique des langues Régionales

  - Acquisition et normalisation de ressources (corpus et lexiques)
  - Développement d'outils pour l'acquisition et l'analyse de corpus
  - Diffusion auprès du grand public

Plan de la présentation

- Langues « peu dotées »
- La langue occitane

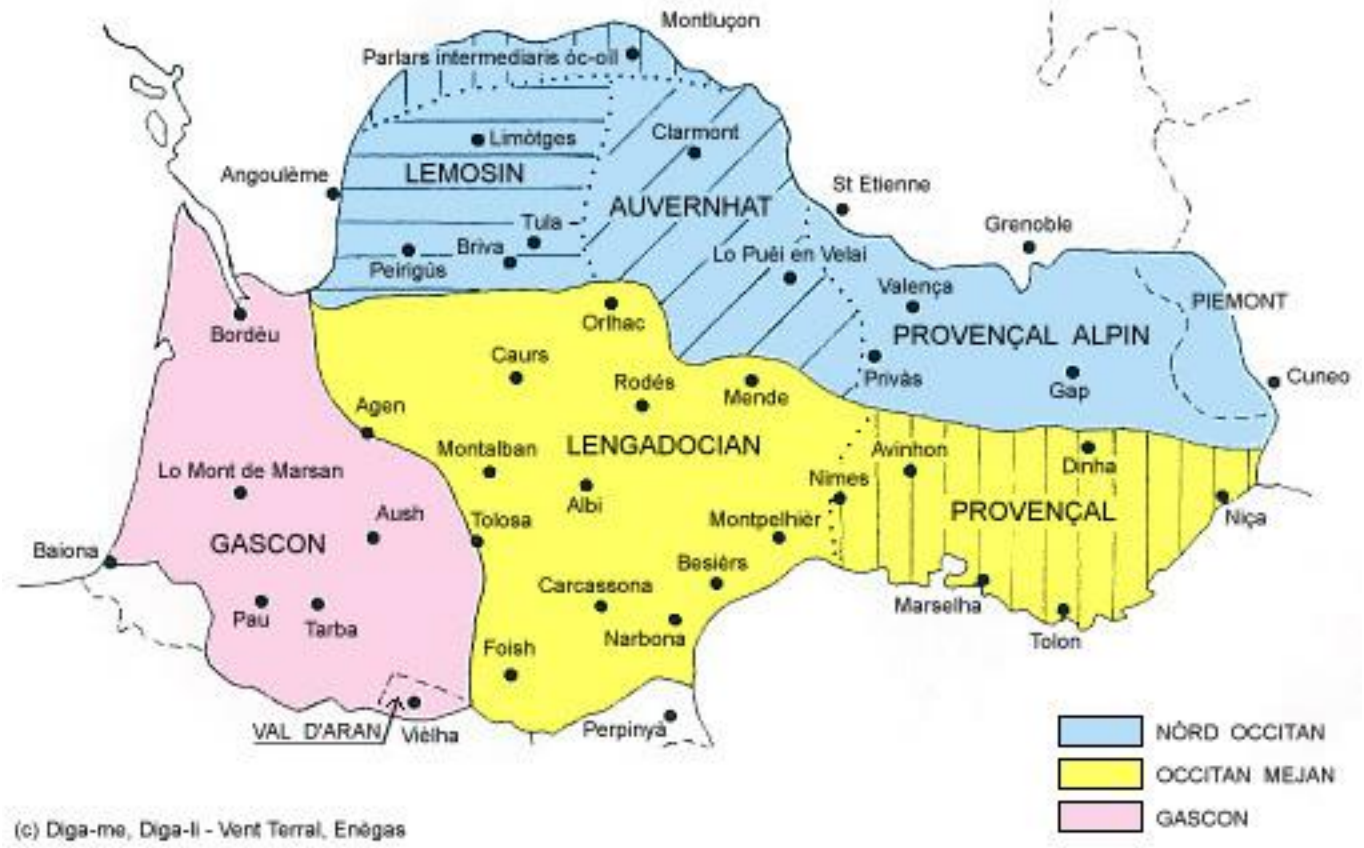
PARTIE I

- Passage de la version expérimentale à la version opérationnelle de BaTelÒc

PARTIE II

- Océrisation avec Jochre
- Analyse morphosyntaxique avec Talismane
- Loflòc

Aire linguistique occitane

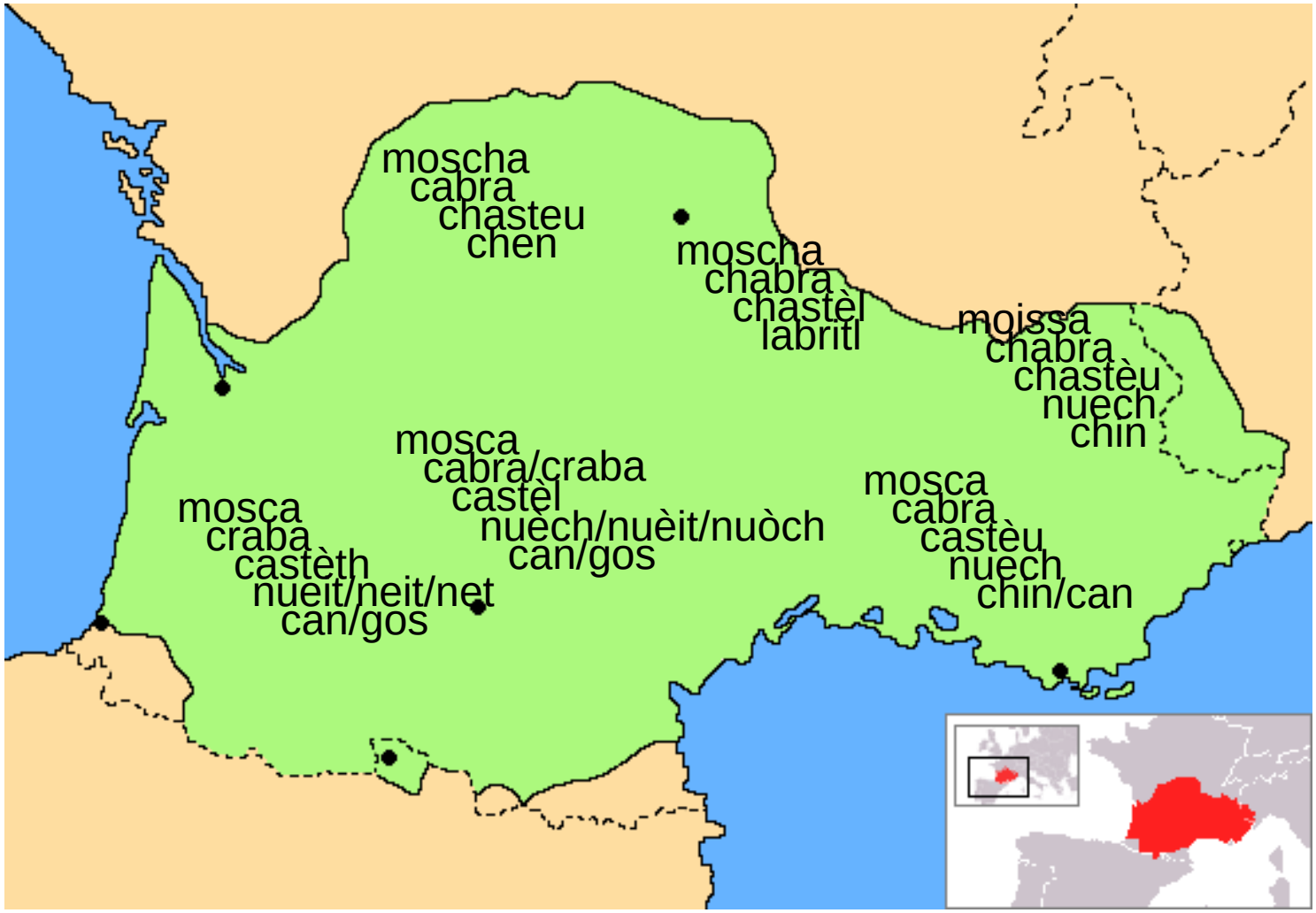


(c) Diga-me, Diga-li - Vent Terral, Enèigas

Langue romane

Français	Occitan	Catalan	Castillan	Italien	Portugais
chanter	**cantar**	**cantar**	**cantar**	**cantare**	**cantar**
je chant**e**	canti	cant**o**	cant**o**	cant**o**	cant**o**
tu chant**es**	cant**as**	cant**es**	cant**as**	canti	cant**as**
il chant**e**	cant**a**	cant**a**	cant**a**	cant**a**	cant**a**
nous chant**ons**	cant**am**	cant**em**	cant**amos**	canti**amo**	cant**amos**
vous chant**ez**	cant**atz**	cant**eu**	cant**áis**	cant**ate**	cant**ais**
ils chant**ent**	cant**an**	cant**en**	cant**an**	cant**ano**	cant**am**
mouche	mosca	mosca	mosca	mosca	mosca
amie	amiga	amiga	amiga	amica	amiga
amour	amor	amor	amor	amore	amor
chèvre	cabra	cabra	cabra	capra	cabra
château	castèl	castell	castillo	castello	castelo
table	taula	taula	mesa	tavolo	mesa

Variation interne



Systèmes graphiques

- Moyen-Age

Cançon de Santa Fe
Chanson de Sainte Foy d'Agen

- XIXème siècle

Canti **lou** Segala, l'**ou**stal que m'a **b**ist naisse.
Pegas**o**, laiss**o**-me, **t'en** podes ana paisse ;
N'aura lèu sieis mil**o** ans que trot**o**s pel pa**b**at :
D**ib**es esse arr**a**ndut, et si**o**s **b**iel acabat.

- XXème siècle

Canti **lo** Segala**r**, l'**o**stal que m'a **v**ist nàisse**r**.
Pegas**a**, laiss**a**-me, **te'n** pòdes ana**r** pàisse**r** ;
N'aurà lèu sièis mil**a** ans que tròt**a**s pel pa**v**at :
D**ev**es èsse**r** arr**e**ndut, e si**á**s **v**ièlh acabat.

Bessou, D'al brès a la toumba

Système de la graphie classique

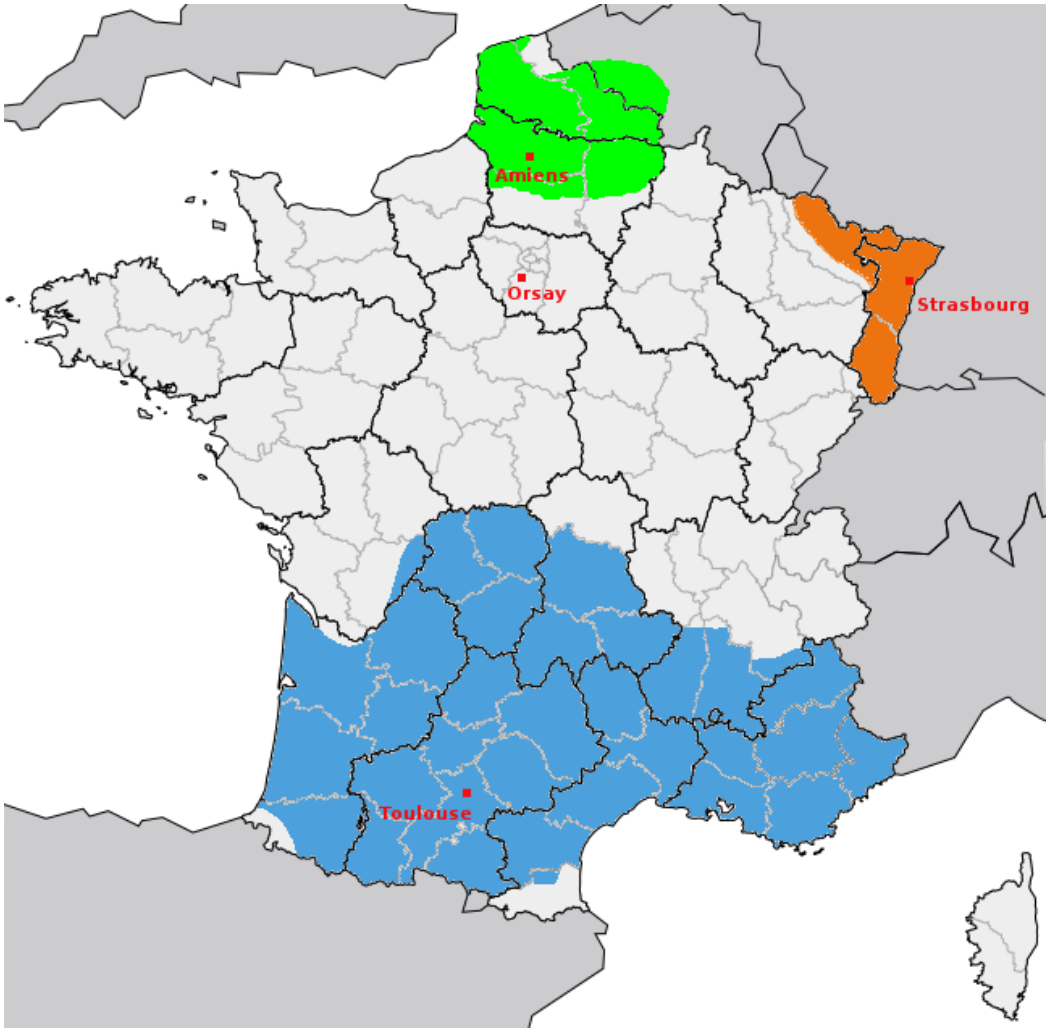
- Lien graphie/phonie non explicite (Sibille, 2002)
  - amo**r** (amour)
  - **v**a**ca, vacas** (vache)
  - n**uèch** (nuit)
  - pri**m** (printemps)
  - assa**j**ar (essayer)
- …
  - pel/peu (cheveux)
  - pèl/pèth/pèu (peau)
  - abstrach / abstrait (abstrait)
  - albèrga / aubèrga (auberge)
  - aparença / aparéncia (apparence)
  - avian / avián (ils avaient)
  - abatre/abattre (abattre)

Parenthèse sur la graphie

- Lien graphie/phonie explicite (ORTHAL)

Français	Alsacien
cuisine	**kuch**, **kucha**, **kische**, **k**hés**che**, **kùch**, **kücha**, **kuche**, **kiche**, **kuchi**
lundi	**mon**dàà, **man**tig, **man**dig, **mon**dàà, **mon**doe, **mon**dàj, **maan**di, **man**di

Retour aux langues « peu dotées »

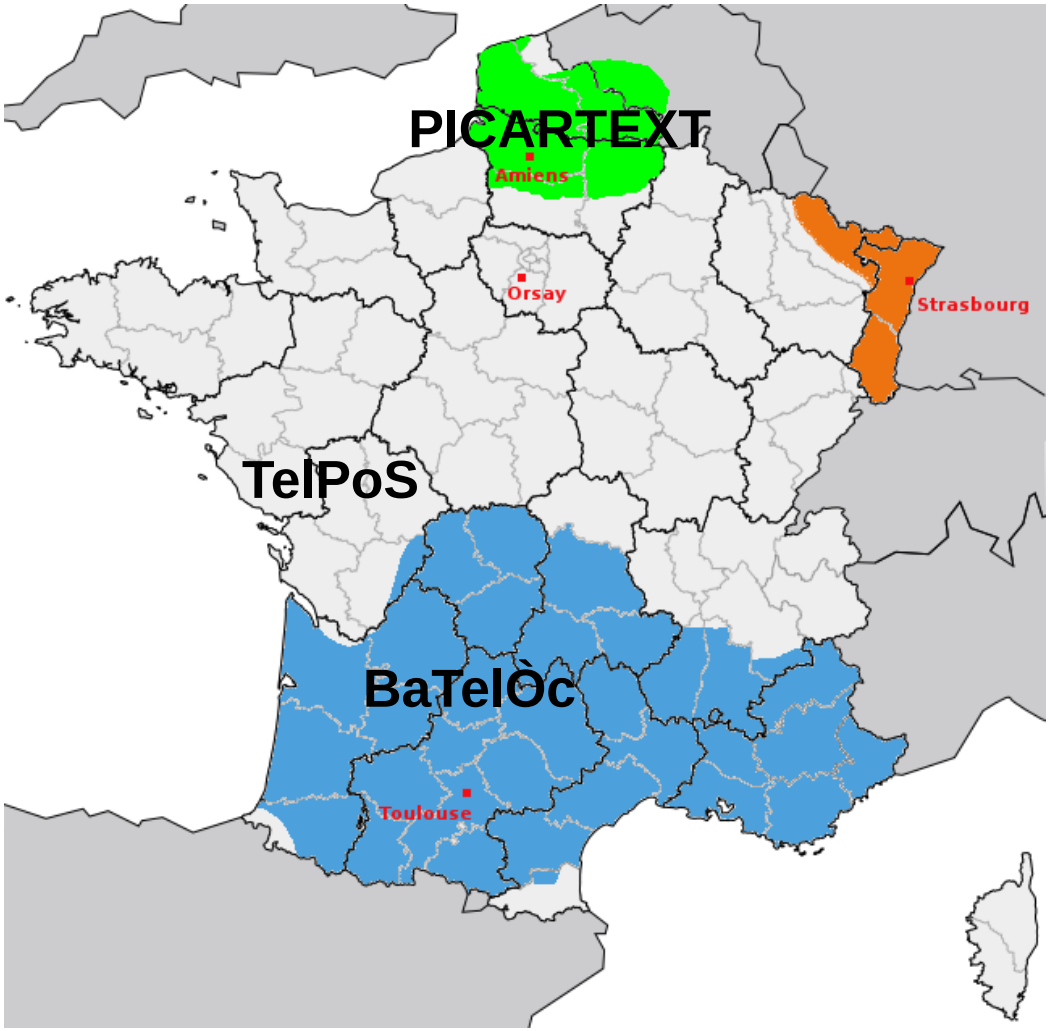


PARTIE I

BaTelÒc – Base de Textes pour la Langue d'Òc

Passage de la base expérimentale à la base opérationnelle

Ressources textuelles



Constitution de corpus (Mosel, 2013)

	Linguistique de corpus traditionnelle		Corpus de documentation des langues
Statut de la langue	Langues “bien étudiées”		Langues “moins étudiées”/Langues en danger
Contenu	Données orales et écrites		Enregistrements Transcriptions Traductions
Ressources recueillies par	Equipes de locuteurs natifs		UN locuteur non natif
Taille	Millions de mots En augmentation		Moins d’UN million de mots

suite

Données disponibles	Grande quantité de données augmentant quotidiennement		Peu de données
Buts	Recherche linguistique (académique et industrielle)		Conservation, linguistique et recherche interdisciplinaire
Compilation	Echantillon représentatif d'une variété à l'étude Constitution d'un corpus de référence		Toutes les opportunités possibles

Constitution de corpus

	Linguistique de corpus traditionnelle	Linguistique de corpus pour les langues de France	Corpus de documentation des langues
Statut de la langue	Langues “bien étudiées”	Langues en danger	Langues “moins étudiées”/Langues en danger
Contenu	Données orales et écrites	Données écrites et orales	Enregistrements Transcriptions Traductions
Ressources recueillies par	Equipes de locuteurs natifs	Equipe de locuteurs non natifs	UN locuteur non natif
Taille	Millions de mots	Plus de UN million de mots	Moins d’UN million de mots

suite

Données disponibles	Grande quantité de données augmentant quotidiennement	Quantité de données importante	Peu de données
Buts	Recherche linguistique (académique et industrielle)	Conservation, linguistique et recherche interdisciplinaire	Conservation, linguistique et recherche interdisciplinaire
Compilation	Echantillon représentatif d'une variété à l'étude Corpus de référence	Toutes les opportunités possibles (certaines variétés étant mieux représentées que d'autres)	Toutes les opportunités possibles

BaTelÒc – base expérimentale

- Accueillir la diversité linguistique
  - 1 million de mots, 35 textes contemporains
  - 3 genres : roman, nouvelles, mémoires
  - 3 dialectes, 1 graphie
- S'aider de l'existant
  - Codage xml TEI P5
  - Se placer dans le respect du droit d'auteur
- S'appuyer sur les compétences du labo
  - Base de données bibliographiques pour métadonnées
  - Prototype de moteur de recherche (recherche plein texte)
  - En ligne accès restreint

Passage à la base opérationnelle

- Augmentation des corpus
  - 3 millions de mots, 90 textes des époques modernes et contemporaines
  - Constitution de sous-corpus
  - 5 dialectes, 3 graphies
  - 9 genres : poésie, essai, traité…
- Interface
  - Segmentation en vue de l'indexation
  - Spécification de l'interface
  - Tests de l'interface
- Autorisation de 4 éditeurs, quelques textes libres de droits

Démonstration de l'interface en 2ème partie

BaTelÒc : **Ba**sa **Te**xtuala per la **l**enga d'**Òc**

[Acuèlh] [Causida del còrpus] [Cèrca simpla] [Cèrca avançada] [Ajuda] [Projècte] [Contacte]

Acuèlh

BaTelÒc es una basa textuala en lenga occitana que recampa d'òbras escrichas de mai d'un genre (roman, teatre, poësia, conte, premsa...) del sègle XIXen a l'ora d'ara e qu'aculhís la variacion grafica e dialectala.

Aqueste site prepausa una interfàcia de consultacion d'òbras occitanas per cercar de formas (mots, partidas de mots, sequéncias de mots) dins un còrpus que podètz definir. L'interfàcia permet pas lo telecargament ni la lectura dels tèxtes complets. [Ne saber mai sul projècte...]

Cercar un mot dins lo còrpus de descobèrta : [**?**]

Cercar

(sensibla a la caissa)

[Se causir un còrpus de trabalh] [**?**]

PARTIE II

TALÒC Traitement automatique de la langue occitane

Motivations

1) OCR

- Ajout de textes pas encore au format numérique
- Besoin d’un OCR de qualité et adapté à la langue dans sa variété

2) Annotations morphosyntaxiques

- Passage de textes nus à des textes annotés
- Lemmatisation et annotations morphosyntaxiques
- Nouvelles modalités de recherche

Construire des outils en TAL pour langues « peu dotées » - Ne pas réinventer la roue

- Regarder ce qui existe déjà/ ce qui se fait pour les langues « très dotées »
- Regarder ce qui existe déjà/ce qui se fait pour la langue concernée
- Transferts technologiques depuis des langues mieux dotées
- Adopter les formats standard

Océrisation de nouveaux textes

- Regarder ce qui existe déjà/ ce qui se fait pour les langues « très dotées »
  - Peu ou pas de recherches car considéré comme un problème résolu
  - Logiciels commerciaux
  - Outils génériques probabilistes (Tesseract, Jochre)
- Regarder ce qui existe déjà/ce qui se fait pour la langue concernée
  - Utiliser des logiciels commerciaux (ABBYY FineReader 12 propose le Provençal ; sans utilisation de lexique ; sans préciser la graphie)
- Transferts technologiques depuis des langues mieux dotées
  - Catalan, Portugais, Allemand
- Adopter les formats standard
  - XML ALTO

Deux outils

Tesseract

- Développé par Google
- Opensource depuis 2005
- Modèles existants : français, allemand, catalan, portugais

- Application de modèles existants via gImageReader (possibilité de combiner plusieurs modèles)
- Entraînement d'un modèle pour l'occitan via jTessBoxEditor

Jochre

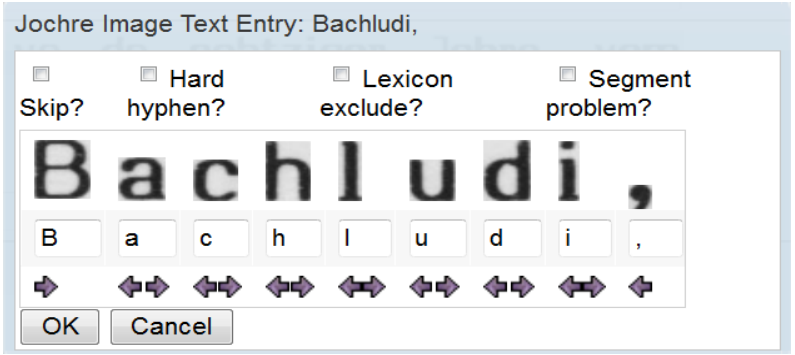
- Développé par Assaf Urieli
- Modèles existants : yiddish

- Entraînement de nouveaux modèles

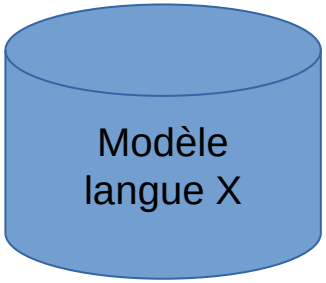
Principes de fonctionnement de l'OCR, exemple de Jochre

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

Mise en correspondance forme - caractère



Apprentissage d'un modèle spécifique à la langue :
→ Extraction de "traits"
→ Utilisation de classifieurs robustes (entropie maximale - MaxEnt ; SVM linéaire)

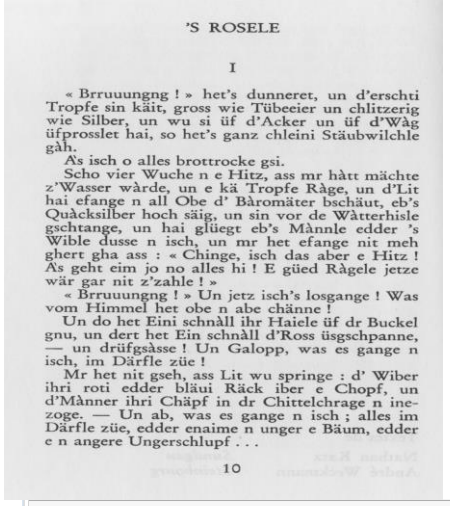


Constitution des corpus d'entraînement

Tesseract

- Entraînement à partir d'images générées
- Faisant varier le type de police (Arial, Times New Roman, Normal et *Italique*)

JochreWeb



'S ROSELE

I

« Brruuungng ! » het's dunneret, un d'erschti Tropfe sin käit, gross wie Tübeeier un chlitzerig wie Silber, un wu si üf d'Acker un üf d'Wàg üfprosslet hai, so het's ganz chleini Stäubwilchle gàh.

Às isch o alles brottrocke gsi.

Scho vier Wuche n e Hitz, ass mr hàtt mächte z'Wasser wàrde, un e kä Tropfe Ràge, un d'Lit hai efange n all Obe d' Bàromäter bschäut, eb's Quàcksilber hoch säig, un sin vor de Wàtterhisle gschtange, un hai glüegt eb's Mànnle edder 's Wible dusse n isch, un mr het efange nit meh ghert gha ass : « Chinge, isch das aber e Hitz ! Às geht eim jo no alles hi ! E güed Ràgele jetze wär gar nit z'zahle ! »

« Brruuungng ! » Un jetz isch's losgange ! Was vom Himmel het obe n abe chänne !

Un do het Eini schnàll ihr Haiele üf dr Buckel gnu, un dert het Ein schnàll d'Ross üsgschpanne, — un drüfgsàsse ! Un Galopp, was es gange n isch, im Därfle züe !

Mr het nit gseh, ass Lit wu springe : d' Wiber ihri roti edder bläui Räck iber e Chopf, un d'Mànner ihri Chäpf in dr Chittelchrage n inezoge. — Un ab, was es gange n isch ; alles im Därfle züe, edder enaime n unger e Bäum, edder e n angere Ungerschlupf . . .

10

Entraînement à partir d'images réelles

Rôle du lexique

Tesseract

- Liste de formes fléchies
  - 1000 mots les plus fréquents
  - Intégralité du lexique
- Système de pondération

Jochre

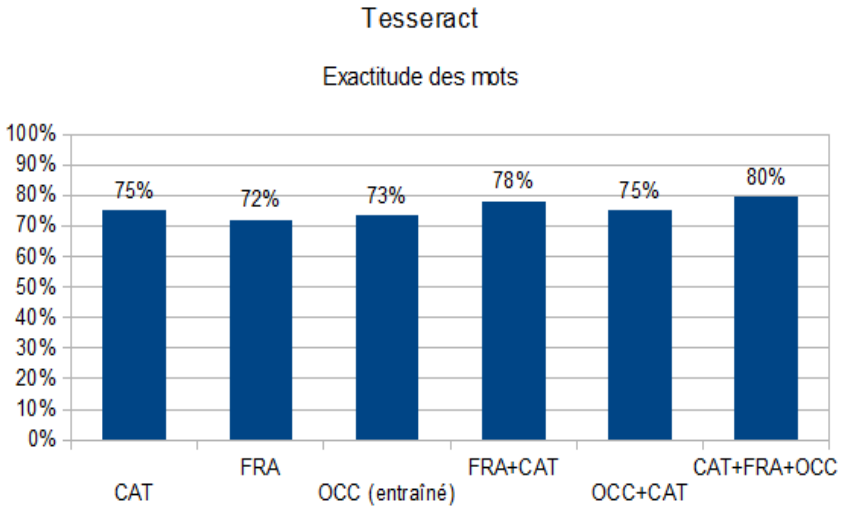
- Liste de formes fléchies
- Système de pondération

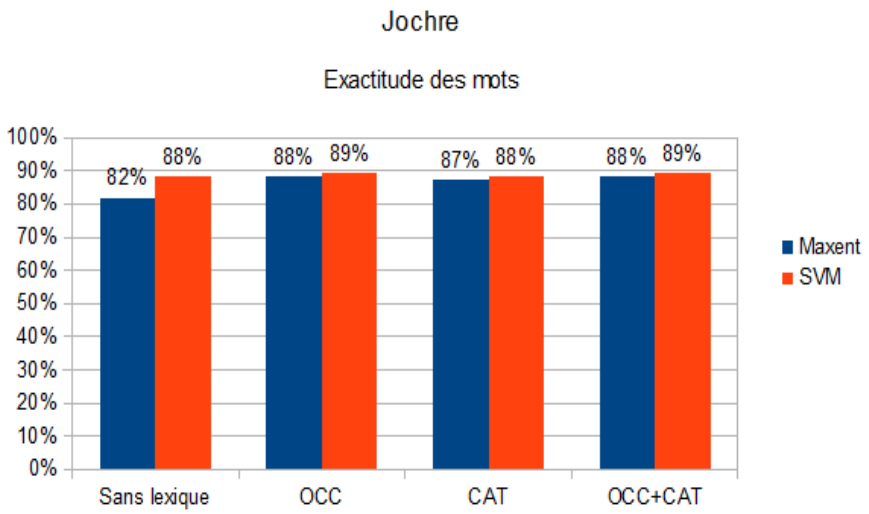
	score initial	connu ?	score ajusté
acordot	**72,0 %**	non (x 0,5)	36,0 %
acordat	70,1 %	**oui** (x 1,0)	**70,1 %**
acordet	64,3 %	non (x 0,5)	32,2 %

Expérience 1 Entraînement d'un modèle Occitan graphie classique

- Corpus (environ 20 400 mots)
  - 11 ouvrages : 7 pour l'entraînement / 4 pour l'évaluation
  - 86 pages
  - 84 000 caractères
  - Années 1960-2000
  - Homogène sérif + (un peu d')italique
- Lexiques occitans (environ 432 000 formes)
  - Lemmes issus de dictionnaires bilingues français/occitan
  - Formes fléchies issues de BaTelÒc (et autres réservoirs de textes)
  - Formes fléchies du conjugueur Verb'òc
- Lexique Catalan d'Apertium (environ 715 000 formes)

Résultats


Tesseract
Exactitude des mots
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
75%
72%
73%
78%
75%
80%
CAT
FRA
OCC (entraîné)
FRA+CAT
OCC+CAT
CAT+FRA+OCC


Jochre
Exactitude des mots
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
82%
88%
88%
89%
87%
88%
88%
89%
Sans lexique
OCC
CAT
OCC+CAT
Maxent
SVM

L’amitié catalano-occitane

Caractères spécifiquement français	Total : 8
Ç	7
œ	1

Caractères spécifiquement catalans	Total : 1238
ò	933
í	118
ó	100
á	68
ú	14
Ò	5

- Les caractères français couvrent **98,53 %** du corpus...

… les caractères catalans : **99,9 %**

Expérience 2 Entraînement d'un modèle Occitan graphies non classiques

- Corpus
  - 18 ouvrages
  - 146 pages
  - 1860-1910
- Lexiques occitans (environ 75 000 formes)
  - Formes fléchies issues de BaTelÒc (et autres réservoirs de textes)

Annotation des corpus dans JochreWeb
Démonstration possible en 2ème partie

Création	Modèle Occitan Graphie Classique	Modèle Occitan Graphies « non classiques »
Modèle	Pas de modèle	Modèle Occitan Graphie Classique
Temps d'annotation (par page)	35 minutes en moyenne	20 minutes en moyenne

La suite dans Jochre Search

- Océrisation de 50 textes du CIRDOC (1860-1910)
- Moteur de recherche sur textes océrisés
- Interface de correction

Démonstration en 2ème partie

Jochre en Occitan

Textes océrisés et lexique par **CLLE-ERSS**
Textes numérisés par **Cirdoc**
Textes indexés by Assaf Urieli, **Joliciel Informatique**

Marianne
Logout

miralh **cercar**

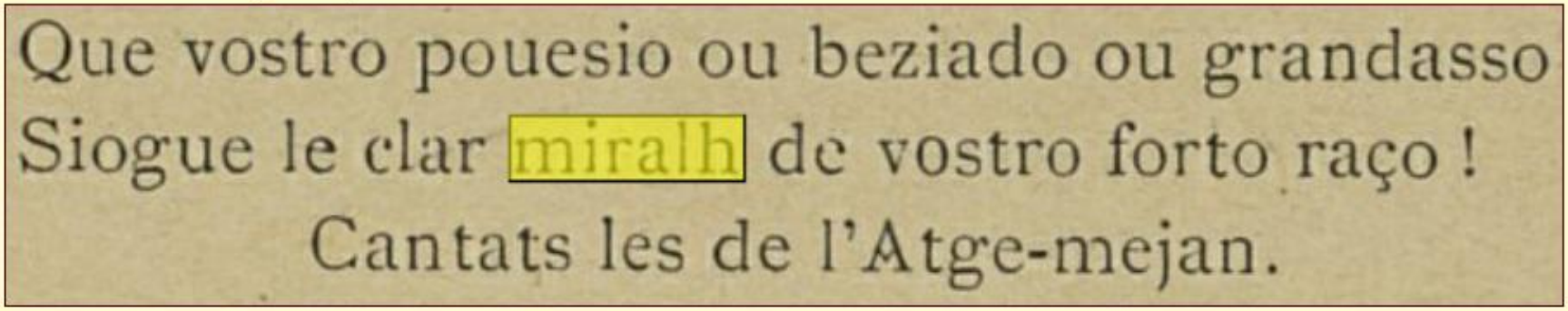
1 resultatas. Resultatas 1 a 1:

Títol: La Muso Silvestro / Auguste Foures ; [notice biographique et litteraire par Gaston Jourdanne]

Fasedor: Fourès¸ Auguste (1848-1891), Jourdanne¸ Gaston (1858-1905). Préfacier¸ etc.

Encors: Paginas 39 a 54

50 Que vostro pouesio ou beziado ou grandasso Siogue le clar **miralh** de vostro forto raço ! Cantats les de l'Atgc-nne,an



Analyse morphosyntaxique

- Regarder ce qui existe déjà/ ce qui se fait pour les langues « très dotées »
  - Outils spécifiques
  - Outils génériques probabilistes
- Regarder ce qui existe déjà/ce qui se fait pour la langue concernée
  - Apertium (traduction automatique pour paires de langues étymologiquement proches : catalan/occitan ; espagnol/occitan)
- Transferts technologiques depuis des langues mieux dotées
  - Catalan ?
- Adopter les formats standard
  - Tagset GRACE

Etape 1 : Test d'Apertium

- Analyse morphosyntaxique = étapes au sein du système global
  - Sélectionne l'étiquette morphosyntaxique la plus probable parmi les étiquettes trouvées dans un lexique
- Extraire les annotations morphosyntaxiques
  - Pas toujours les mêmes unités de segmentation (patron de traduction)
  - Pas d'annotations pour tous les mots
  - Lexique avec deux variantes (impact sur les annotations ?)
  - Pas de processus d'évaluation (pour cette étape)
- Amélioration d'Apertium
  - Augmentation/Séparation des lexiques
  - Révision de l'ensemble des étiquettes morphosyntaxiques

Etape 2 : Outils génériques

- Constituer des corpus annotés et des lexiques
  - 1) Traiter chaque dialecte/parler comme une langue
    - Avantages : ressources spécifiques de très grande qualité
    - Désavantages : couteux
  - 2) Exploiter les similarités pour passer d'un dialecte/parler à l'autre
    - Avantages : moins couteux
    - Désavantages : ressources de moins bonne qualité
- 2) Commencer par un dialecte « Languedocien » en construisant un corpus de texte homogène au sein du parler « Rouergue »
  - Etendre au dialecte languedocien
  - Etendre à plusieurs dialecte
- Amorce avec Apertium

Corpus

	Source	Variété	Taille (mots)
Corpus d'entraînement	*E la barta floriguèt* (Molin)	UN seul auteur Languedocien - Rouergue	~ 2500
Corpus d'évaluation	*Los crocants de Roergue* (Delèris) *Dels camins bartassièrs* (Esquieu) *Hont blanc* (Lavit)	Languedocien - Rouergue Languedocien - L-et-G Gascon - Bigorre	~ 500 ~ 500 ~ 500
Lexique	*Dictionnaire Français/Occitan Languedocien* (Laux/Congrès) Une partie du Verb'Òc (Sauzet/Congrès)	Languedocien	~ 200 000

Exemple d’annotations

Index	Forme	Lemme	Cat. principale	Morphologie
12	Canti	cantar	Vm	i-p1s-
13	lo	lo	Da	-ms-d
14	Segalar	Segalar	Np	??
15	,	,	F	
16	l'	lo	Da	-ms-d
17	ostal	ostal	Nc	ms
18	que	que	Pr	-ms--
19	m'	me	Pp	1ms?-
20	a	a	Sp	
21	vist	veire	Vm	p-s-sm
22	nàisser	nàisser	Vm	n-----
23	.	.	F	

Tagset GRACE

- 1er niveau : 10 catégories (compatible Eagles)

N (nom), V (verbe), A (adjectif), P (pronom), D (déterminant), R (adverbe), S (préposition), C (conjonction), I (interjection), X (résidu)

- 2ème niveau : 32 catégories (compatible Grace)

Nc, Np, Vm, Va, Af, Ao, Ak, As, Pp, Pi, Ps, Pt, Pr, Px, Pk, Da, Dd, Ds, Di, Dt, Dk, Dp, Rg, Rx, Rp, Rq, Sp, Sd, Cc, Cs, I, X

- 3ème niveau : 153 catégories (spécifique Occitan)

N	**c**	**ms**
P	**p**	**2msa-**

Talismane

- Analyseur morphosyntaxique et syntaxique par apprentissage supervisé
- Logiciel libre
- http://redac.univ-tlse2.fr/talismane.html
- Exactitude de 97% en français et en anglais

Entraîner Talismane – Expé 1

- Annoter morphosyntaxiquement deux dialectes occitans avec les ressources d'un seul dialecte
  - Corpus d'entraînement Languedocien Rouergat
  - Lexique Languedocien
  - 3 Corpus d'évaluation (Languedocien Rouergat, Languedocien (Lot et Garonne), Gascon)

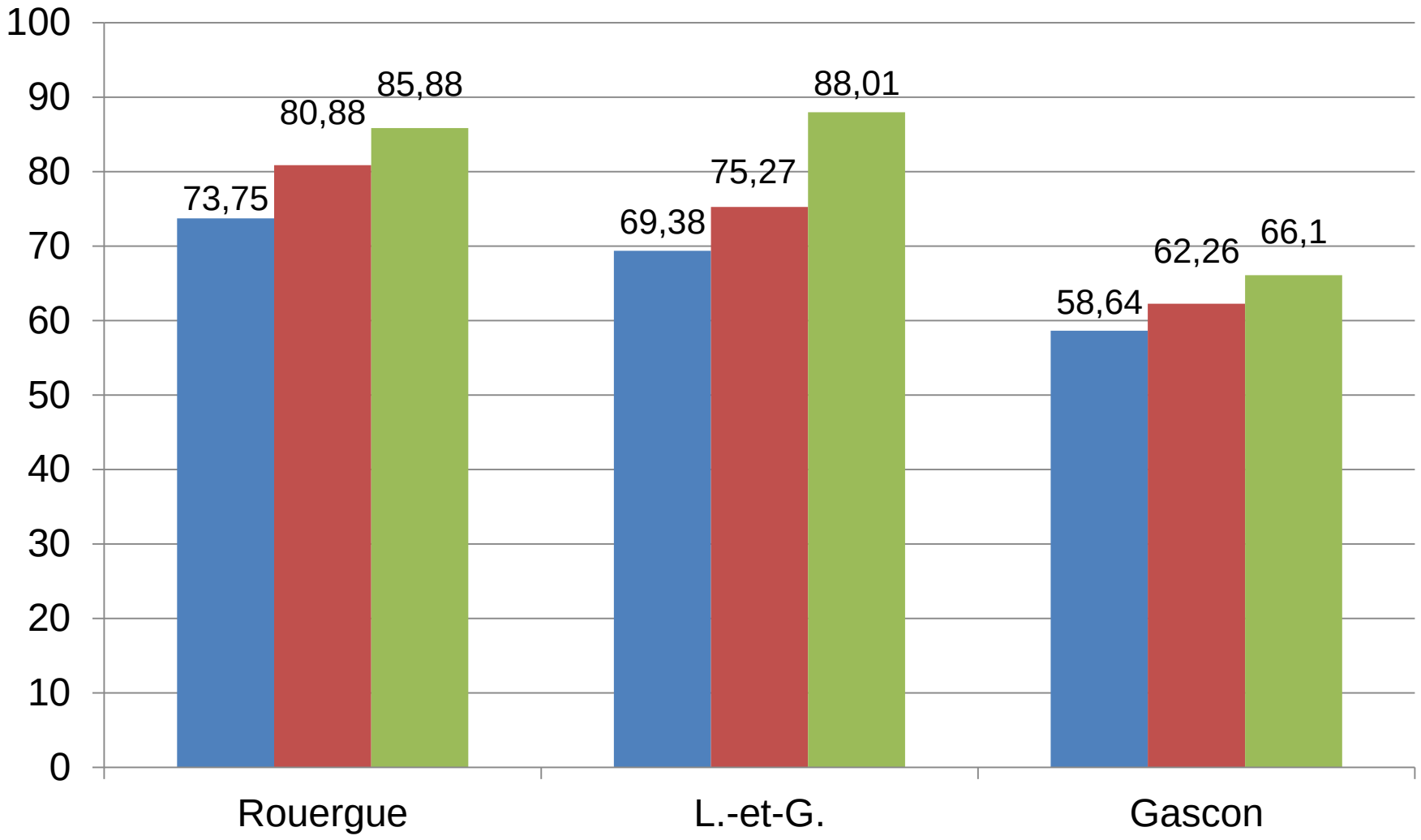
- Vaut-il mieux constituer des lexiques ou des corpus annotés ?

Couverture des lexiques

Corpus	Entraînement	Rouergue	Lot-et-Garonne	Gascon
Taille	2 501	701	467	469
Inconnu dans le lexique	0,1%	16,6%	19,9%	40,1%
Inconnu dans le lexique (classes ouvertes)	0,2%	29%	37,3%	59,1%
Inconnu dans le lexique (classes fermées)	0,0%	1,5%	1,1%	20,4%

Résultats

Sans lexique
Avec lexique de classes fermées
+ lexique de classes ouvertes



Entraînement Corpus vs. Lexique

- Rappel des ressources :
  - Corpus : 2 500 mots
  - Lexique : 200 000 mots
- Expériences avec division par deux des ressources
  - 2 corpus de 1 250 mots
  - 2 lexiques de 200 000 mots
- Gains
  - En doublant le corpus d'entraînement : 1,46%
  - En doublant le lexique : 4,16%

Entraîner Talismane – Expé 2

- Améliorer les résultats pour le languedocien
  - Petit corpus d'entraînement Languedocien
  - Gros corpus d'entraînement Catalan (AnCora (500 000 mots))

Couverture des lexiques

Corpus	Entraînement Languedocien	Entraînement Catalan	Evaluation Languedocien
Taille	2 800	500 000	1 200
Inconnu dans le lexique	2,4%	39,4%	19,9%
Inconnu dans le lexique (classes ouvertes)	4,3%	62,8%	37,3%
Inconnu dans le lexique (classes fermées)	0,1%	12,5%	1,1%

Expériences et Résultats

- Petit Corpus Occitan Languedocien *vs.* Gros Corpus Catalan

	Occitan (2 500)	Catalan (500 000)
Exactitude	89,04	90,11

- Transposition Catalan *vs.* Occitan des 250 mots les plus fréquents

	Catalan transposé
Exactitude	91,10

- Combinaison des 2 corpus

	Catalan transposé + Occitan (x 1)	Catalan transposé + Occitan (x200)
Exactitude	91,26	92,26

Conclusions

- Talismane pour l'occitan (paramétrage)
- Résultat acceptable avec peu de données (89-92%)
- Importance de construire des lexiques à large couverture
- Mise au point du manuel d'annotation
- Utile d'utiliser les ressources des langues étymologiquement proches
- Très prometteur pour l'analyse inter-variantes

Augmentation des corpus

Nom	Taille (lignes)
Mouly_Barta	8 749
Bodon_Drac	4 207
Marti_Miramont	6 402
Deleris_Crocants	2 725
Bessou_Tomba	5 173
Esquieu_Bartassiers	512
TOTAL	27 768

Loflòc

- **L**exic **o**bert de **f**ormas **fl**echidas de l'**oc**citan
- Accueillir par étapes toute la **variation** (et de faire des liens ?)
- Utiliser des formats standards

Source	Taille (mots)
Dictionnaire Français/Occitan Languedocien (Laux, 2005)	50734
Dictionnaire Occitan/Français Languedocien (Laux, 2001)	43391
Ajout de lexique des classes fermées	1993
Génération des pluriels pour Laux FR/OC	42911
Génération des pluriels pour Laux OC/FR	32727
Verb'Òc Languedocien (Sauzet & Ubaud, 2005 ; Sauzet, 2016)	646464
TOTAL (cumulé)	828230

En chiffres

	Nombre de lemmes	Nombre de formes fléchies
Nom	35460	68 135
Verbe	13024	645 151
Adjectif	10885	43 261
Pronom	161	405
Déterminant	72	163
Adverbe	1386	1 405
Préposition	580	1 047
Conjonction	113	177
Interjection	194	194
TOTAL	**61 878**	**759 941**

Variation sur les formes fléchies

Forme fléchie	Lemme	Morphologie
soi	ésser	Présent de l'indicatif 1ère sg
ès siás	ésser	Présent de l'indicatif 2ème sg
es	ésser	Présent de l'indicatif 3ème sg
sèm	ésser	Présent de l'indicatif 1ère pl
sètz	ésser	Présent de l'indicatif 2ème pl
son	ésser	Présent de l'indicatif 3ème pl

Variation sur les lemmes

- Variation dialectale et intradialectale sur les lemmes

Lemmes	Formes au pluriel	Id	Source
cabra	cabras	23	Laux (languedocien)
cabra	cabras	23	Basic gasc/lang
craba	crabas	23	Basic gasc/lang
craba	crabas	23	Per Noste (gascon)
chabra	chabras	23	Faure (Vivaro-alpin)

- 23 = {{cabra}, {cabra, craba}, {craba}, {chabra}}

Conclusion

- BaTelÒc - Accès au grand public
- Jochre, Talismane, Loflòc - Accès en public restreint

- Méthodes par apprentissage supervisé
- Constitution des ressources étapes par étapes

Perspectives

- Individual Fellowship (Queens University, Belfast avec Janice Carruthers)
  - Constitution d'un corpus de contes occitan
  - Annotations de traits linguistiques temporels et discursifs (temps verbaux, connecteurs, cadres de discours)
  - Description linguistique des structures de discours (différents degrés d'oralité, comparaison français/occitan…)
- Projet Poctefa Linguapyr (déposé)
  - Annotations syntaxiques d'un corpus occitan
  - Parser pour l'occitan (Talismane)