ACTES DU XII e/n CONGRÈS DEL

de l'Associacion Internacionala d'Estudis Occitans

de l'Association Internationale d'Études Occitanes

ALBI, 10-15/07 2017

édités par Jean-François Courouau
en collaboration avec David Fabié

editats per Joan-Francés Courouau
en collaboracion amb Dàvid Fabié

Fidelitats et dissidences

élités et dissidences

Section française de l'Association internationale d'études occitanes

SFAIEO

Vol. 1

Myriam BRAS
Nabil HATHOUT
Jean SIBILLE
Université Toulouse – Jean Jaurès
CLLE-ERSS

Marianne VERGEZ-COURET
Université de Poitiers
FoReLLIS

Aure SÉGUIER
Benaset DAZÉAS
Lo Congrès permanent de la lenga occitana

Loflòc : Lexic obèrt flechit occitan

1. Un lexique informatisé des formes fléchies de l'occitan

Loflòc (Lexic obèrt flechit occitan - Lexique ouvert fléchi occitan) est un lexique informatisé de formes fléchies en occitan. Ses premières versions ont été réalisées dans le cadre du projet ANR RESTAURE[1] (Bernhard et Vergez-Couret, 2016 ; Bernhard et al. 2018) puis du projet européen POCTEFA LINGUATEC[2], en collaboration avec Lo Congrès Permanent de la Lenga Occitana[3].

La création d'un lexique informatisé pour l'occitan s'intègre dans un projet plus global de création de ressources linguistiques informatisées pour une langue qui dispose de peu de ressources à l'heure actuelle. Ces ressources, qu'elles soient lexicales comme Loflòc, ou textuelles comme BaTelÒc[4] (Bras et Thomas 2011, Bras et Vergez Couret 2016), sont conçues en suivant un double objectif : d'une part la préservation et la diffusion du patrimoine linguistique et d'autre part la création de ressources pour le développement d'outils de traitement automatique des langues (analyseurs morpho-syntaxiques, analyseurs syntaxiques, traduction automatique). La création de ces ressources permet de compléter certaines étapes définies par la *Feuille de route pour le développement du numérique occitan* (Lo Congrès Permanent de la Lenga Occitana, 2014 ; Dazéas, 2015, Séguier et Mercadier, 2016), et de développer des applications de recherche et d'extraction d'information, des agents conversationnels, des outils d'aide à l'écriture comme les claviers prédictifs, des correcteurs orthographiques….

Les objectifs qui ont présidé à la création de Loflòc sont les suivants :

- Doter l'occitan d'un lexique structuré de formes fléchies adapté aux besoins du TAL (Traitement Automatique des Langues) et d'une ressource indispensable pour les modules de base du traitement automatique : lemmatiseur, analyseur morphosyntaxique, analyseur syntaxique (Vergez-Couret et Urieli, 2015) ;
- Munir le lexique d'une interface de consultation ;
- Utiliser un jeu d'étiquettes morphosyntaxiques standard ;
- Accueillir par étapes la variation (dialectale, intra-dialectale, graphique).

Nous décrivons ici la première version de Loflòc, telle que présentée au congrès de l'AIEO 2017, tout en intégrant les évolutions concernant la catégorisation de certains items permises par

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

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

[3] http://www.locongres.org

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

le travail d'annotation morphosyntaxique de corpus mené dans les projets RESTAURE et LINGUATEC. Nous présenterons dans la section 2 les sources lexicographiques et flexionnelles que nous avons utilisées pour construire Loflòc, puis nous décrirons en section 3 les informations morphosyntaxiques associées aux formes fléchies qu'il regroupe. Dans la section 4, nous résumerons la méthodologie suivie et nous indiquerons, en section 5, la structure des fichiers et des éléments quantitatifs. Enfin, nous expliquerons notre stratégie pour la prise en compte de la variation en section 6, et nous terminerons par une présentation de l'interface de consultation de Loflòc.

2. Sources lexicographiques et flexionnelles

Pour constituer le lexique Loflòc, la stratégie choisie a été d'intégrer en premier lieu des ressources disponibles au format numérique, de les enrichir avec des informations grammaticales lorsque ces dernières sont incomplètes ou inadaptées, et de compléter les paradigmes flexionnels (genre et nombre…). Les premières ressources lexicographiques intégrées à Loflòc pour le languedocien sont les parties occitanes des dictionnaires bilingues Occitan-Français et Français-Occitan Languedocien de Laux (2001, 2005). De ces dictionnaires, normalisés au format XML (norme TEI P5) par le Congrès Permanent de la Lenga Occitana[5] pour la mise en ligne via l'application *Dicod'Òc*[6], il a été possible d'extraire automatiquement les lemmes, certaines flexions et les informations catégorielles. Nous avons dans un second temps complété les flexions manquantes (cf. section 4).

Certaines catégories de la grammaire traditionnelle choisies par Laux ont été modifiées pour placer les ressources dans une perspective linguistique. Ainsi par exemple, plusieurs éléments catégorisés comme des adjectifs (indéfinis ou possessifs) ont-ils été recatégorisés comme des déterminants.

Pour la flexion verbale, nous avons bénéficié des données de l'application *vèrb'Òc*[7], conjugueur édité par le Congrès Permanent de la Lenga Occitana à partir de données publiées dans deux ouvrages de conjugaison occitane (Sauzet, 2016 ; Sauzet et Ubaud, 1995). D'autres sources lexicographiques pour le languedocien et les autres dialectes de l'occitan seront exploitées par la suite.

Nous présenterons dans cette section les étiquettes choisies pour les catégories grammaticales (ou parties du discours ; *Part Of Speech, POS*), et indiquerons les origines de ce jeu d'étiquettes, puis nous décrirons chaque catégorie, en commençant par les catégories formant la partie ouverte du lexique (noms, adjectifs, verbes, et une partie de la classe des adverbes) et en continuant avec les catégories formant sa partie fermée (déterminants, prépositions, pronoms, conjonctions, adverbes).

3. Informations morphosyntaxiques associées aux formes fléchies

Loflòc est construit sur le modèle des lexiques français existants tels que Morphalou (Romary et al., 2004) et GlaFF (Sajous et al., 2013).

Nous avons adopté les étiquettes du standard GRACE (Rajman et al., 1997) lui-même adapté du jeu d'étiquettes MULTEXT (Ide & Véronis, 1994) et EAGLES (von Reckowski, 1996). Nous avons choisi de laisser les étiquettes en anglais tout en les adaptant aux spécificités de l'occitan afin de faciliter la comparaison de notre lexique aux lexiques des langues proches qui ont également adopté des jeux d'étiquettes semblables et comparables (français, catalan).

[5] http://www.locongres.org/images/docs/choix_encodage.pdf
[6] https://www.locongres.org/fr/applications/dicodoc-fr/
[7] https://www.locongres.org/fr/applications/2014-02…/verboc-recherche

3.1 Trois niveaux d'étiquettes

Les étiquettes GRACE comportent 3 niveaux. Le premier niveau indique la catégorie grammaticale des formes fléchies (POS), et sert à identifier les symboles de ponctuation (F) et les formes attestées n'entrant pas dans les catégories précédentes (X).

Etiqueta	Etiquette	Tag	
Nom	Nom	(Noun)	N
Vèrbe	Verbe	(Verb)	V
Pronom	Pronom	(Pronoun)	P
Adjectiu	Adjectif	(Adjective)	A
Determinant	Déterminant	(Determiner)	D
Advèrbi	Adverbe	(Adverb)	R
Adposicion	Adposition	(Adposition)	S
Conjonccion	Conjonction	(Conjunction)	C
Interjeccion	Interjection	(Interjection)	I
Residú	Résidu	(Residual)	X
Ponctuacion	Ponctuation	(Punctuation)	F

Tableau 1. Etiquettes du premier niveau

Le deuxième niveau, appelé Type, propose une classification sémantique ou fonctionnelle spécifique à chaque catégorie de niveau 1.

Les étiquettes suivantes appartiennent au troisième niveau et concernent les informations morphosyntaxiques relatives à la flexion en genre, nombre, personne, temps, mode, etc. Quand un trait n'est pas pertinent, il est indiqué avec le symbole tiret '-'. Les informations manquantes sont signalées par un point d'interrogation '?'.

Les modifications apportées au jeu d'étiquettes GRACE sont signalées au fur et à mesure de la description du jeu d'étiquettes ci-dessous.

3.2 Les noms (N)

Les étiquettes des noms sont les mêmes que celle de GRACE pour le français. Elles sont présentées dans le tableau 2. En plus de l'étiquette N au premier niveau, il y a une étiquette par sous-catégorie : par exemple, la forme *abelhas* sera étiquetée « Ncfp » : nom commun féminin pluriel.

	Nom (V)	Valeur	Code	Exemple
Niv 1			N	
Niv 2	Type	common proper cardinal	c p k	abelha Pèire tres
Niv 3	Genre	masculine feminine	m f	libre abelha
	Nombre	singular plural	s p	abelha abelhas

Tableau 2. Etiquettes des noms

3.3 Les verbes (V)

Les étiquettes des verbes sont présentées dans le tableau 3. Par exemple, la forme *manjariam* est étiquetée « Vmc-p1p- ». Le jeu d'étiquettes choisi pour le français dans GRACE a été enrichi pour rendre compte de la différence de forme entre l'impératif positif et l'impératif négatif (1, 2) :

(1) Manja !
(2) Manges pas !

Pour encoder cette spécificité de l'occitan, nous avons ajouté l'attribut « Form » qui peut prendre deux valeurs « positive » ou « negative ». Ce trait est seulement pertinent pour la valeur « imperative » de l'attribut « Mood ».

	Verbe	Valeur	Code	Exemple
Niv 1			V	
Niv 2	Type	main	m	manjar
		auxiliary	a	aver
Niv 3	Mood/Vform	indicative	i	mangi
		subjunctive	s	mange
		imperative	m	manja
		conditional	c	manjariái
		infinitive	n	manjar
		participle	p	manjat
	Form	positive	a	manja
		negative	n	manges
	Tense	present	p	mangi
		imperfect	i	manjavi
		future	f	manjarai
		past	s	mangèri
	Person	first	1	mangi
		second	2	manjas
		third	3	manja
	Number	singular	s	mangi
		plural	p	manjam
	Gender	masculine	m	manjat
		feminine	f	manjada

Tableau 3. Etiquettes des verbes

3.4 Les adjectifs (A)

Le tableau 4 présente les étiquettes des adjectifs qui comportent 5 types : la forme *polidas* sera par exemple étiquetée « Afpfp » (adjectif qualificatif positif féminin pluriel).

La valeur « Degree » n'est pertinente que pour les adjectifs qualificatifs. Les adjectifs comparatifs de l'occitan sont *màger* (3), *melhor* (4), *mendre* et *pièger* :

(3) lo problèma màger es que sabèm pas ont son passats
(4) lo moment melhor es quand lo solelh trescòla

Les comparatifs peuvent être variables ou invariables selon les parlers. Il sont systématiquement annotés aux 3 niveaux.

Nous reviendrons sur les adjectifs ordinaux, cardinaux et possessifs en section 3.10.

Niv 1	Adjectif (A)	Valeur	Code	Exemple
Niv 2	Type	qualificative ordinal cardinal indefinite possessive	f o k i s	bon centen dos autre meu
Niv 3	Degree	positive comparative	p c	bon melhor
	Genre	masculine feminine	m f	bon bona
	Nombre	singular plural	s p	bon bons

Tableau 4. Etiquettes des adjectifs

3.5 Les adverbes (R)

Le tableau 5 présente les étiquettes des adverbes. Au deuxième niveau, on distingue les adverbes généraux, les particules, les adverbes interrogatifs et exclamatifs, les adverbes intensifs et quantificatifs. Cette dernière sous-catégorie est ajoutée au modèle GRACE pour des adverbes tels que *plan, tant* … qui ont la particularité de pouvoir dans certains parlers s'accorder en genre et en nombre (5) :

(5) de pomas, n'i a planas

Niv 1	Adverbe (R)	Valeur	Code	Exemple
Niv 2	Type	general particle interrogative/exclamative intensive/quantitative	g p t q	aisidament ne quant plan
Niv 3	Degree	positive comparative negative	p c n	aisidament melhor pas
	Genre	masculine feminine	m f	plan plana
	Nombre	singular plural	s p	plana planas

Tableau 5. Etiquettes des adverbes

La catégorie des adverbes généraux (Rg) regroupe la partie ouverte de la catégorie des adverbes, constituée par les adverbes en *–ment*. Elle comporte également une liste fermée d'adverbes comme *uèi, ara, puèi,* et les adverbes négatifs comme *pas* ou *non* quand ils sont utilisés seuls pour marquer la négation.

Les adverbes *ne* et *non* sont analysés comme des particules (Rp) dans les parlers et les situations où la particule accompagne l'adverbe général négatif *pas*. Quand il est employé seul *non* est, comme *pas*, adverbe général négatif (Rgn). D'autres adverbes négatifs, par exemple *jamai, sonque*, sont souvent combinés avec l'adverbe négatif *pas*.

La catégorie des particules (Rp) sera également utilisée pour les particules énonciatives du gascon *que, be, e, ja, si*.

Les adverbes intensifs (Rq) sont *plan, mai, força, cap, pus, brica, mai, tot, tròp, gaire, aitant, un pauc, …*

Les formes adverbiales comparatives sont par exemple *melhor, mens* (6, 7) :

(6) lo vin de l'an passat es melhor que lo vin d'ongan, va melhor, canta melhor que sa sòrre
(7) de noses, ongan, n'avèm mensas

Elles ont la particularité de ne pas pouvoir se combiner avec *mai* (8) :

(8) *mai melhor, *mai mens

3.6 Les pronoms (P)

Le tableau 6 présente les étiquettes des pronoms. Pour le cas, l'étiquette « case » ne s'applique qu'aux pronoms clitiques. On reconnaîtra donc les pronoms toniques par la présence d'un « - » en position 5. Le cas « nominative » est très rare en occitan, puisque, dans la majorité des parlers, il n'y a pas de pronom sujet, mais, cette possibilité existant dans certains parlers périphériques (Briançonnais, Limousin), nous avons gardé ce type. Par ailleurs, nous avons séparé le cas « datif » des deux autres cas obliques du modèle GRACE. Ainsi, « oblique » est remplacé par un cas « datif » et une étiquette « others » qui regroupe le génitif et l'ablatif.

Niv 1	Pronom (P)	Valeur	Code	Exemple
Niv 2	Type	personal	p	ieu
		demonstrative	d	aquò
		indefinite	i	mantuns
		possessive	s	meu
		interrogative	t	quin
		relative	r	qui
		reflexive	x	se
		cardinal	k	dos
Niv 3	Person	first	1	ieu
		second	2	tu
		third	3	el
	Genre	masculine	m	quin
		feminine	f	quina
	Nombre	singular	s	quin
		plural	p	quines
	Case	nominative	n	ieu
		accusative	a	lo, o, ne
		dative	d	li
		others	o	ne
	Possessor	singular	s	meu
		plural	p	nòstre

Tableau 6. Etiquettes des pronoms

Au sujet des pronoms personnels, il est intéressant de mentionner le fait que les sources lexicographiques suggèrent l'existence d'un genre neutre pour des pronoms pouvant renvoyer soit à des référents inanimés (*aquò, aiçò*), soit à des faits ou à des événements (*o, zo, ac, ba,*

aquò…). En réalité, il n'y a pas de flexion spécifique à ce genre potentiellement « neutre », ces pronoms s'accordant au masculin. Il n'y a donc pas lieu d'introduire un genre « neutre ».

Enfin, les pronoms indéfinis couvrent les cas des identificateurs (9) et des quantificateurs non cardinaux (10) :

(9) D'unes cresián vertadièrament a una galejada (Viaule, BaTelÒc)
(10) Mantuns avián fachas tres o quatre sasons (Delèris, BaTelÒc)

Voici un exemple d'étiquetage des deux pronoms clitiques de la phrase *lo li balhi* :

- *lo* est étiqueté « Pp3msa-» (accusatif)
- *li* est étiqueté « Pp3-sd-» (datif)

Le pronom réflexif *me* sera étiqueté « Px1-s-- ». Le pronom *aquò* sera étiqueté « Pd-ms-- ». Les pronoms personnels clitiques réalisant des anaphores propositionnelles comme *o*, *zo*, *ac*, *ba* seront étiquetés « Pp3msa- ».

3.7 Les déterminants (D)

Le tableau 7 présente les étiquettes des déterminants.

Niv 1	Déterminants (D)	Valeur	Code	Exemple
Niv 2	Type	article	a	lo
		demonstrative	d	aquel
		possessive	s	mon
		indefinite	i	cada
		inter./exclam.	t	quin
		relative	r	lo qual
		cardinal	k	dos
		partitive	p	de
Niv 3	Person	first	1	mon
		second	2	ton
		third	3	son
	Genre	masculine	m	mon
		feminine	f	ma
	Nombre	singular	s	mon
		plural	p	mos
	Possessor	singular	s	mon
		plural	p	nòstre
	Nature	definite	d	lo
		indefinite	i	un

Tableau 7. Etiquettes des déterminants

Les déterminants articles (Da) sont constitués par la série *lo, la, los, las, un, una, de.*

Les déterminants indéfinis (Di) sont *cada*, *qualque*, *mantun*, *mantuns*, *mai d'un*, *tot*, *un pauc de*… (11) :

(11) **mantuns** còps, **cada** jorn, **tota** la vila

Nous donnons ci-dessous un exemple de déterminant relatif (Dr) (12) :

(12) qu'aimava asagada d'un veiròt de riquiquí (***lo qual*** veiròt demorava al fons del bòc) (Escafit, BaTelòc)

Par rapport au jeu d'étiquettes GRACE, l'attribut « partitive » a été ajouté au trait « Type ». Il existe en effet un déterminant partitif simple, *de*, étiqueté « Dp-ms- - », par exemple en (13) :

(13) que vòl far amb **de** sucre (Landièr, BaTelòc)

Pour les autres formes possibles (dans les dialectes du nord), *del*, *de la,* deux étiquettes seront sollicitées, Dp+Da. Par exemple *de la* en (14) sera étiqueté « Dp-fs-- + Da-fs-d » :

(14) A la poncha d'un pueg, un faus, e **de la** mossa (Roux, BaTelòc)

Enfin, on peut noter que la forme *un* peut être étiquetée « Da-ms-i » (déterminant article indéfini) ou « Dk-ms - - » (déterminant cardinal). EAGLES propose arbitrairement de toujours choisir le type « Article ». Nous choisissons d'intégrer dans le lexique les deux codes. La désambiguïsation pourra être faite lors de l'annotation des corpus en fonction du contexte. En cas de doute, nous donnerons la préférence au type « Article ».

3.8 Les prépositions (S)

Niv 1	Préposition (S)	Valeur	Code	Exemple
Niv 2	Type	preposition deictic	p d	dins vaquí

Tableau 8. Etiquettes des prépositions

La catégorie S regroupe les prépositions (Sp) : *per, de, coma, dins, abans, dempuèi, a, sus, jós, en*… et les déictiques (Sd) comme *vaquí.*

Les amalgames d'une préposition et d'un déterminant (*del, dels, al, als, pel, pels, sul, suls, jol, jols,*…) sont codés avec les deux étiquettes correspondant à la forme non amalgamée (Sp+Da). Par exemple *del*, étant équivalent à *de lo (de le)* sera codé *« Sp+Da-ms—d »*.

3.9 Les conjonctions (C)

Niv 1	Conjonction (C)	Valeur	Code	Exemple
Niv 2	Type	coordinating subordinating	c s	e que

Tableau 9. Etiquettes des conjonctions

La catégorie C regroupe les conjonctions de coordination (Cc) *mas, e, o* … et les conjonctions de subordination (Cs) *quand, coma, que, se* ….

3.10 Étiquettes des cardinaux, des ordinaux et des possessifs

Nous avons introduit dans les sections 3.2, 3.4, 3.6 et 3.7 les catégories Nk, Ak, Pk et Dk pour les noms, les adjectifs, les pronoms et les déterminants cardinaux, dont nous illustrons ci-dessous les emplois en contexte (15-17) :

(15) un parelh de dos (*dos* est Nk)
(16) los dos amics (*dos* est Ak)
(17) Cinc ostals dins lo Causse (…). Dos son barrats (Gairal, BaTelÒc) (*dos* est Pk, *cinc* est Dk)

De même, nous distinguons les numéraux ordinaux selon qu'ils sont adjectifs (18) ou pronoms (19) :

(18) lo tresen còp (*tresen* est Ao)
(19) lo tresen (*tresen* est Po)

Nous classons également les possessifs selon leur catégorie grammaticale : adjectif possessif As (20), déterminant possessif Ds (21) et pronom possessif Ps (22) :

(20) lo meu libre / lo libre meu

(21) mon libre
(22) aquel libre es lo meu / aquel libre es meu

GRACE diffère sur ce point d'autres standards comme EAGLES ou Universal Dependencies (Nivre et. al 2016) qui classent tous les cardinaux et tous les possessifs dans une même catégorie, facilitant ainsi le classement des formes numérales, mais ne permettant pas une catégorisation qui est possible et pertinente pour l'analyse morphosyntaxique et syntaxique.

3.11 Autres étiquettes

Trois autres étiquettes font également partie du jeu d'étiquette de Loflòc. Il s'agit :
- des interjections (I) : *rai, òu, ai, zo, i, a, o, òu, flica-flaca, pam …*
- des résidus (X), que nous utilisons pour noter les voyelles épenthétiques : *-n-*, *-s-* …
- de la ponctuation (F) : . ; , - ! ….

4. Méthodologie de construction de Loflòc

4.1 Extraction des formes à partir des sources lexicographiques

Les formes ont été extraites ou générées à partir des entrées de dictionnaires bilingues Occitan/Français et des traductions des dictionnaires bilingues Français/Occitan. Toutes les variantes présentes dans ces ressources ont été conservées. Le premier traitement a consisté à extraire de ces ressources les unités lexicales (généralement données sous leur forme de citation dans les dictionnaires ; en TAL, les formes de citation sont appelées lemmes), les formes fléchies (le cas échéant) et les variantes (le cas échéant) ainsi que les informations morphosyntaxiques. Nous détaillons ci-dessous le procédé spécifique à chacune de nos trois sources.

Dictionnaire français/occitan de C. Laux

Nous sommes partis d'une ressource au format XML construite à partir du fichier doc fourni par l'éditeur. Pour chaque entrée en français, nous avons extrait la (ou les) traductions ainsi que toutes les informations morphosyntaxiques pertinentes (les informations en gras dans l'exemple ci-dessous) :

```
<entry n="10">
<form><orth>abandonner</orth><gramGrp><pos
norm="verb">v</pos></gram></gramGrp></form>
<sense n="I-A-1-a">
<cit type="translation" xml:lang="oc-lnc">
<form><orth>abandonar</orth><gramGrp><iType
type="altvb">o</iType></gramGrp></form>
</cit>
<cit type="translation" xml:lang="oc-
lnc"><form><orth>daissar</orth></form></cit></sense></entry>
```

Partir d'un dictionnaire dont les entrées sont en français a posé quelques difficultés. Certaines traductions proposées ne pouvaient pas être intégrées telles quelles dans le lexique : nous avons manuellement supprimé les formes composées décomposables comme *cabana clujada (cabane équipée d'un toit de chaume)* proposée en traduction de *chaumière* ; les paraphrases proposées quand il n'existe pas de traduction directe comme pour *crochetteur : cambriolaire que fòrça las pòrtas (cambrioleur qui force les portes)*. Nous avons également vérifié et corrigé manuellement toutes les formes répertoriées à la fois comme nom commun masculin et féminin. Enfin, nous avons vérifié manuellement la pertinence de toutes les traductions données au pluriel pour déterminer si la forme proposée dans la glose pouvait être utilisée comme lemme ou pas (par exemple la forme *asmòlhas (grosses pinces ou tenailles de forgeron)* est choisie comme lemme mais la forme *preparatius (préparatifs)* doit être associée au lemme *preparatiu)*.

Dictionnaire occitan/français de C. Laux

Nous sommes partis d'une base de données au format SQL développée par Lo Congrès à partir de la ressource lexicographique. De cette base de données ont été extraites toutes les entrées en occitan ainsi que les informations morphosyntaxiques associées. Toutes les formes nouvelles (par rapport à la liste des formes extraites du dictionnaire français/occitan) ont été validées manuellement (correction de petites coquilles dues à divers traitements informatiques). Les catégories grammaticales manquantes de 200 formes ont été renseignées manuellement.

Conjugueur VerbÒc du Congrès

La ressource a été construite par Lo Congrès et se présente sous la forme d'un fichier csv contenant les champs : lemme, modèle de flexion selon Sauzet et Ubaud (1995), forme fléchie, informations morphosyntaxiques et étiquette EAGLES.

Contrairement aux choix faits dans Verbòc, nous avons décidé de ne conserver que les 3es personnes du singulier pour les verbes impersonnels et défectifs : *caler, faler, replòure, plòure, nevar.*

4.2 Complétion des étiquettes et des paradigmes

Complétion des étiquettes (lexiques des classes fermées)

Les lexiques des classes fermées (déterminants, pronoms, adverbes quantifieurs, adjectifs possessifs, locutions prépositionnelles) ont été complétés manuellement à partir des introductions des dictionnaires de Laux mais également de grammaires telles que Taupiac (2008).

Féminins et pluriels des noms et des adjectifs

Nous avons reconstruit les lemmes et les formes de certains noms communs féminins et les formes féminines de certains adjectifs. Toutes ces formes ont été générées à partir des informations données dans les dictionnaires (23) :

(23) menteur, -euse adj/n *messorguièr, -a, mentidor, -oira, menteire, -a*.

Nous avons ensuite effectué la génération des pluriels des noms et des adjectifs, sauf pour les pluriels irréguliers qui étaient signalés dans les dictionnaires. Nous avons pour cela compilé un ensemble de règles dans un programme Perl. Pour une finale donnée, nous avons défini une ou plusieurs règles spécifiant des cas particuliers et une règle générale qui s'applique par défaut.

Par exemple, pour un nom ou un adjectif se terminant par -g, les règles suivantes s'appliquent dans l'ordre :

		Règle appliquée	Exemple
1	[séqu. de lettres][consonne]g	Ajout de s	dramaturg/dramaturgs
2	[séqu. de lettres] [LÒG|FAG|PAG|TÈG|FUG]	Ajout de s	epilòg/epilògs
3	castig, escag, grog	Ajout de s	escag/escags
4	[séqu. de lettres]g	Ajout de es	puèg/puèges

Tableau 10. Règles de flexion du pluriel pour les noms et les adjectifs en *-g*

Une fois toutes les formes fléchies produites, nous avons effectué une vérification des paradigmes. Dans la majorité des cas, les noms ont un paradigme de deux formes : singulier et pluriel tandis que les adjectifs ont un paradigme de quatre formes : masculin singulier, féminin singulier, masculin pluriel et féminin pluriel. Nous avons manuellement vérifié les paradigmes défectifs (forme en moins ou en plus), ce qui nous a permis de corriger des coquilles dans les formes extraites des dictionnaires et dans les règles de formation des pluriels.

4.3 Fusion et référencement des sources

Nous avons fusionné les deux lexiques puis extrait toutes les formes identiques avec deux étiquettes une au masculin et une au féminin (par exemple *poèma* codé Nom commun féminin

dans le dictionnaire occitan/français de C. Laux (LAUXOCFR) et Nom commun masculin dans le dictionnaire français/occitan de C. Laux (LAUXFROC). Quand les deux formes ne sont pas des homographes, comme pour *poèma*, seule l'étiquette correcte, ici Nom commun masculin, sera retenue. Nous avons manuellement vérifié tous les cas en nous référant au DOGMO (Ubaud, 2011).

Référencement des sources

La constitution de Loflòc a fait intervenir diverses ressources, deux dictionnaires et verb'Òc mais à terme nous utiliserons également d'autres ressources lexicographiques et grammaticales et des corpus de textes. Nous avons fait le choix de conserver dans le lexique la provenance de chaque élément intégré. Cela constitue également un moyen de renseigner la forme du point de vue dialectal.

Sources	Codes de référence	Dialecte
Dictionnaire Occitan-Français Languedocien de Laux (2001)	LAUXOCFR	Languedocien
Dictionnaire Français-Occitan Languedocien de Laux (2005)	LAUXFROC	Languedocien
Formes fléchies plurielles des noms et adjectifs de LAUXFROC	Loflòc via LAUXFROC	Languedocien
Formes fléchies plurielles des noms et adjectifs de LAUXOCFR	Loflòc via LAUXOCFR	Languedocien
Verb'òc	VERBOC	Languedocien

Tableau 11. Sources de Loflòc

5. Structure et taille de Loflòc

5.1 Structure du lexique

Loflòc est une table excel où chaque ligne comporte 6 champs[8] :

Identifiant (Id)	Forme Fléchie (FF)	Lemme (L)	Etiquette de FF (TFF)	Etiquette de L (TL)	Source

Tableau 12. Structure de Loflòc

En voici un extrait à titre d'illustration :

147	abdicar	abdicar	Vmn-----	Vmn-----	LauxFROC
148	abdomèn	abdomèn	Ncms	Ncms	LauxFROC
149	abdominal	abdominal	Afpms	Afpms	LauxFROC
150	abdominala	abdominal	Afpfs	Afpms	LauxFROC
151	abduccion	abduccion	Ncfs	Ncfs	LauxFROC

Tableau 13. Extrait de Loflòc

Chaque forme fléchie possède un identifiant unique. Les identifiants sont attribués de façon chronologique au moment de l'intégration de chaque nouvelle ressource. Avant intégration, les formes fléchies de la nouvelle ressource sont préalablement ordonnées alphabétiquement.

[8] Dans les versions ultérieures de Loflòc, nous avons indiqué séparément la source du lemme et la source de la flexion, ajoutant ainsi un champ.

5.2 Taille du lexique

Dans la version décrite ici Loflòc contient 759 938 formes fléchies pour 61 875 lemmes et 257 étiquettes.

LauxFROC	50 734
LauxOCFR	43 391
Loflòc	1 993
Loflòc via LauxFROC	42 911
Loflòc via LauxOCFR	32 727
Verb'Òc Languedocien	646 464
Total cumulé (nombre de lignes)	828 230
Total sans les doublons	759 938

Tableau 14. Nombre d'entrées par source

	Nombre de lemmes	Nombre de formes fléchies
Nom	35 460	68 135
Verbe	13 024	645 151
Adjectif	10 885	43 261
Pronom	161	405
Déterminant	72	163
Adverbe	1 386	1 405
Préposition	580	1 047
Conjonction	113	177
Interjection	194	194
Total	61 875	759 938

Tableau 15. Nombre de lemmes et de formes fléchies par POS

6. Gestion de la variation

Les variations, qu'elles soient dialectales, intradialectales ou graphiques, sont présentes dans les productions en occitan, anciennes et actuelles. Les outils automatiques, tout comme les locuteurs (locuteurs par transmission familiale, néo-locuteurs, apprenants…), sont confrontés à toutes ces variations. Afin de bâtir des outils les plus robustes possibles, cette variation doit être décrite et représentée dans les lexiques. En outre, dans les outils de consultation et d'interrogation du lexique, l'utilisateur pourra découvrir et mieux appréhender toute la variation possible.

Comme nous l'avons dit en introduction, nous souhaitons accueillir dans le lexique toutes les variations dialectales, intradialectales et graphiques. Autant que possible, nous souhaitons également ne pas nous contenter de lister les formes variantes dans une liste plate mais nous voulons les lier entre elles. L'objectif est de constituer des ensembles de lexèmes, comme par exemple pour le mot *chevreuil* traduit dans trois dictionnaires par : {{cabiròl, cabròl}$_1$, {cabròu}$_2$, {cabiròu}$_3$}. Le premier ensemble contient les traductions issues de Laux FR/OC$_1$ (languedocien), puis nous avons une traduction issue du CREO$_2$ (provençal) et enfin une traduction issue de Per Noste$_3$ (gascon). Ces variantes sont intradialectales pour le contenu du sous-ensemble 1, et elles sont interdialectales pour les 3 sous-ensembles.

600 paires de variantes ont été extraites du Dictionnaire Français/Occitan de Laux (2005). Ce sont uniquement des cas de variantes intradialectales. Les variantes sont des paires de formes de citation. Nous listons ces variantes dans un fichier dont la structure ainsi qu'un extrait sont présentés dans le tableau 16. Le tableau 17 présente la description des formes fléchies correspondantes dans Loflòc.

Id_FF_1	Id_FF_2	S	FF_1	FF_2	T
identifiant de la forme fléchie 1	*identifiant de la forme fléchie 2*	*Source*	*forme fléchie 1*	*forme fléchie 2*	*Etiquette*
345	348	LauxFROC	abrial	abril	Ncms
409	410	LauxFROC	absténer	abstenir	Vmn-----
96728	96726	Loflòc via LAUXFROC	abrials	abrils	Ncmp

Tableau 16. Structure du fichier variantes et extrait

Id	FF	L	T_FF	T_L	Source
345	abrial	abrial	Ncms	Ncms	LauxFROC
348	abril	abril	Ncms	Ncms	LauxFROC
409	absténer	absténer	Vmn-----	Vmn-----	LauxFROC
410	abstenir	abstenir	Vmn-----	Vmn-----	LauxFROC

Tableau 17. Extrait de Loflòc

7. Navigation et interrogation de Loflòc

Nous avons développé un prototype d'interface d'interrogation de Loflòc (en partant du principe qu'une forme donnée par l'utilisateur est toujours une forme fléchie). L'interface proposera à terme :

- Un outil de lemmatisation de formes fléchies : à partir d'une forme fléchie donnée par l'utilisateur, l'outil renvoie l'ensemble des lemmes possibles ;
- Un moteur de recherche sur les formes fléchies : à partir d'une séquence de lettres, le moteur renvoie l'ensemble des formes fléchies possibles avec les fonctionnalités *est, commence par, finit par, contient* et le choix de la catégorie grammaticale, par exemple les noms qui se terminent par « on » ;
- Un outil de consultation des flexions : à partir d'un lemme sélectionné par l'utilisateur, l'outil renvoie toutes les formes fléchies possibles ;
- Un outil de gestion des variantes : à partir d'un lemme sélectionné par l'utilisateur, l'outil renvoie toutes les variantes listées dans le lexique.

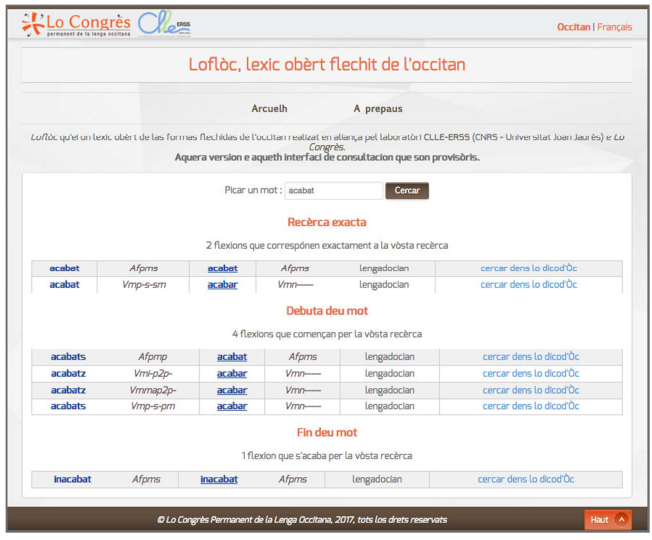

Figure 1. Copie d'écran de l'interface d'interrogation

8. Perspectives

La version de Loflòc présentée ici sera suivie d'autres versions correspondant à autant d'étapes d'enrichissement du lexique :

- Intégration de nouvelles ressources lexicographiques pour le languedocien (partie languedocienne du dictionnaire « Lo Basic » en ligne sur le site du Congrès, dictionnaire de Joan de Cantalausa en ligne sur le site du Congrès) ;
- Intégration de ressources lexicographiques pour le gascon à partir du Basic dans un premier temps pour permettre le lien entre les deux dialectes puis intégration du dictionnaire de *Per Noste* et du Verb'Òc gascon, tous en ligne sur le site du Congrès ;
- Enrichissement de Loflòc à partir de formes nouvelles extraites de la base textuelle BaTelòc : repérage des formes, des lemmes et complétion des paradigmes ;
- Intégration progressive des autres dialectes de l'occitan.

Nous mènerons en parallèle le travail annoncé en section 6 sur les ensembles de lemmes (sortes de supralemmes ou lemset) pour la gestion de la variation.

Références bibliographiques

Références primaires : sources lexicographiques

LAUX, Christian (2001). *Dictionnaire occitan-français. Languedocien*, avec la collab. de Serge Granier, Puylaurens, IEO, Section du Tarn.

LAUX, Christian (2005). *Dictionnaire Français-Occitan*, Castres, IEO Tarn.

SAUZET, Patrick / UBAUD, Josiane (1995). *Le verbe occitan. Lo vèrb occitan*, Aix-en-Provence, Édisud.

SAUZET, Patrick (2016). *Conjugaison occitane*, IEO edicions.

TAUPIAC, Jacme (2008). *Gramatica Occitana*, Institut d'Estudis Occitans.

UBAUD, Josiane (2011). *Diccionari ortografic, gramatical e morfologic de l'occitan*, Canet, Trabucaire.

Références secondaires : articles scientifiques et rapports

BERNHARD, Delphine / VERGEZ-COURET, Marianne (2016). « Le projet RESTAURE », *Les technologies pour les langues régionales de France*, Condé-sur-Noireau, DGLFLF, 96-100.

BERNHARD, Delphine / LIGOZAT, Anne-Laure / MARTIN, Fanny / BRAS, Myriam / MAGISTRY, Pierre / VERGEZ-COURET, Marianne / STEIBLÉ, Lucie / ERHART, Pascale / HATHOUT, Nabil / HUCK, Dominique / REY, Christophe / REYNÉS, Philippe / ROSSET, Sophie / SIBILLE, Jean / LAVERGNE, Thomas (2018). « Corpora with Part-of-Speech Annotations for Three Regional Languages of France: Alsatian, Occitan and Picard », *11th Edition of the Language Resources and Evaluation Conference (LREC). May 2018. Miyazaki, Japan*.

BRAS, Myriam / THOMAS, Joan (2011). « BaTelÒc : cap a una basa informatizada de tèxtes occitans », *in* Angelica Rieger / Domenge Sumien (éds), *Occitània convidada d'Euregio. Lièja 1981 - Aquisgran 2008. Bilanç e amiras. Actes du Neuvième Congrès International de l'Association Internationale d'Études Occitanes, Aix-la-Chapelle, 24-31 août 2008*, Aachen, Shaker, 661-669.

BRAS, Myriam / VERGEZ-COURET, Marianne (2016). « BaTelÒc: A text Base for the Occitan Language », *in* Vera Ferreira / Peter Bouda (eds.), *Language Documentation and Conservation in Europe*, Honolulu, University of Hawai'i Press, 133-149.

DAZÉAS, Benoît (2015). « Feuille de route pour le développement numérique occitan », *Actes du Workshop TALARE (Traitement Automatique des Langues Régionales de France et d'Europe). Caen, juin 2006*.

IDE, Nancy / VÉRONIS, Jean (1994). « MULTEXT (Multilingual Text Tools and Corpora) », *Proceedings of the 14th International Conference on Computational Linguistics, COLING'94, Kyoto, Japan, 1994*.

Lo Congrès (2014). « Diagnostic e huelha de rota tau desvolopament numeric de la lenga occitana 2015-2019. Rapòrt finau deu projècte », *Lo Congrès permanent de la lenga occitana*.

NIVRE, Joakim / DE MARNEFFE, Marie-Catherine / GINTER, Filip / GOLDBERG, Yoav / HAJIČ, Jan Manning, Christopher / McDONALD, Ryan / PETROV, Slav / PYYSALO, Sampo / SILVEIRA, Natalia / TSARFATY, Reut / ZEMAN, Daniel (2016). « Universal Dependencies v1 : A Multilingual Treebank Collection », *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*.

RAJMAN, Martin / LECOMTE, Josette / PAROUBEK, Patrick (1997). Format de description lexicale pour le français. Partie 2 : Description morpho-syntaxique. *Technical report, EPFL & INaLF. GRACE GTR-3-2.1.* https://publi.limsi.fr/tlp/grace/

ROMARY, Laurent / SALMON-ALT, Susanne / FRANCOPOULO, Gil (2004). « Standards going concrete : from LMF to Morphalou. Workshop on Electronic Dictionaries », *The 20th International Conference on Computational Linguistics - COLING 2004*, coling, 2004, Genève.

SAJOUS, Franck / HATHOUT, Nabil / CALDERONE, Basilio (2013). « GLÀFF, un Gros Lexique À tout Faire du Français », *Actes de la conférence Traitement Automatique des Langues Naturelles (TALN 2013), 17-21 juin 2013, Les Sables d'Olonne*.

SÉGUIER, Aure / MERCADIER, Gilbert (2016). « Le numérique au service de la transmission de la langue occitane : situation et perspectives de développement », *Les technologies pour les langues régionales de France*, Condé-sur-Noireau, DGLFLF, 82-90.

VERGEZ-COURET, Marianne / URIELI, Assaf (2015). « Analyse morphosyntaxique de l'occitan languedocien : l'amitié entre un petit languedocien et un gros catalan », *Actes du Workshop Traitement Automatique des Langues Régionales de France et d'Europe, Caen*.

VON REKOWSKY, Ursula (1996). « ELM-FR: A typed French incarnation of the EAGLES-TS – Definition of Lexical Specification and Classification Guidelines », GSI-Erli.