# Prise en compte de la variation dans l'annotation automatique morphosyntaxique de l'occitan

Clamença Poujade[1, 2]
(1) CLLE, UMR 5263, Université de Toulouse Jean Jaurès, CNRS
5, allées Antonio Machado, 31078 cedex 9 Toulouse, France
(2) JOLICIEL Informatique, 2 Av. du Cardie, 09000 Foix, France
clamenca.poujade@univ-tlse2.fr

RÉSUMÉ

L'occitan est une langue romane de France, d'une petite partie de l'Italie et de l'Espagne. Il comprend de nombreuses variations à l'écrit, notamment les variations dialectale et de graphie. C'est un enjeu important dans la dotation de la langue que de pouvoir prendre en compte la variation. Le traitement automatique de l'occitan est en développement cette dernière dizaine d'années. Des ressources et des outils sont constitués et commencent à prendre en compte la variation dialectale. Toutefois, la variation graphique est peu présente dans ces travaux. Notre travail de recherche se concentre sur l'annotation automatique en lemmes, en parties du discours et en flexion verbale d'un corpus de textes contenant ces deux types de variation. À partir de ce corpus nous entraînons des outils d'annotation automatique robustes sur la variation globale de l'occitan.

ABSTRACT

**Variation in Automatic Annotation of Occitan.**

Occitan is a Romance language of France, a little part of Italy and Spain. It includes many written variations, dialectal and spelling variations. Being able to take variation into account is a major challenge to provide the language. Automatic processing of Occitan has been developing over the last ten years. Resources and tools have been developed and are beginning to take dialectal variation into account in these works. However, graphical variation is rarely taken into account. Our research focuses on the automatic annotation into lemmas, parts of speech and verbal inflection of a corpus of texts containing these two types of variation. From this corpus we train robust automatic annotation tools on global variation in Occitan.

MOTS-CLÉS : Annotation automatique - Variation - Langue moins dotée - occitan - parties du discours - corpus.

KEYWORDS: Automatic Annotation - Variation - Less-ressourced Language - Occitan - parts of speech - corpus.


FIGURE 1 – Carte des dialectes de l'occitan (Bernhard *et al.*, 2021)

L'occitan est une langue parlée dans le sud de la France, dans une vallée des Pyrénées en Espagne et dans certaines vallées du Piémont en Italie. Il compte six grands dialectes (Bec, 1995) (Figure 1) et ces dialectes sont constitués de nombreux parlers. La distinction entre les dialectes se fait via des faisceaux d'isoglosses créant un continuum linguistique.

## 1 Ressources et outils pour la linguistique outillée de l'occitan, état des lieux

L'occitan a longtemps été considéré comme une langue peu dotée dans le traitement automatique des langues. Toutefois, depuis une dizaine d'années, la langue occitane s'est dotée de plusieurs corpus numériques et de corpus annotés, d'outils numériques et automatiques de traitement de la langue. Notre groupe de recherche sur l'outillage des langues peu dotées [1] travaille à sa dotation, et un organisme associatif de régulation de la langue produit des applications pour le grand public [2].

Parmi ces ressources, l'occitan dispose d'une base textuelle, BaTelÒc (Bras & Vergez-Couret, 2016) constituant un corpus de presque quatre millions de mots, mis en ligne et interrogeable via une interface. Notre groupe de recherche a également construit un corpus annoté en Parties du discours (POS) dans le cadre du projet ANR Restaure (Bernhard *et al.*, 2018) en collaboration avec d'autres langues de France et un corpus annoté en POS et en dépendances syntaxiques (Tolosa Treebank) dans le cadre du projet européen Linguatec, avec d'autres langues des Pyrénées (Miletic *et al.*, 2020a). En comparaison avec des langues bien dotées, ces corpus sont de petite taille, ils comptent seulement quelques dizaines de milliers de mots. Ces deux corpus annotés ont pris en compte une certaine variation linguistique avec la présence de quatre dialectes dans le corpus Restaure et de cinq dialectes dans le corpus Linguatec (Miletic *et al.*, 2020b). En collaboration avec Lo Congrès, nous avons également construit un lexique de formes fléchies (Bras *et al.*, 2020) qui a servi à l'entraînement de premiers outils d'annotation automatique pour l'occitan. Les outils

1. Groupe OCRE : https ://clle.univ-tlse2.fr/accueil/equipes-de-recherche/sciences-du-langage/occitan-langues-romanes-langues-deurope-decrire-formaliser-outiller-comparer
2. Lo Congrès Permanent de la lenga occitana : https ://locongres.org/

entraînés avec ces ressources fournissent de bons résultats pour l'annotation en parties du discours et dépendances syntaxiques des différents dialectes ; mais ces ressources ne contiennent pas de variation graphique.

## 2 Enjeux du traitement de la variation

L'occitan est une langue n'ayant pas de standard vraiment défini. Beaucoup de textes sont écrits dans un parler qui est propre à l'auteur ou l'autrice.

La graphie non plus n'a pas de norme générale arrêtée. Cependant, plusieurs normes se font concurrence : la graphie "classique" (Exemple 1), qui est celle construite avec l'objectif d'avoir une graphie propre à la langue (et qui est celle présente dans les corpus annotés occitans déjà construits) ; la graphie dite "mistralienne" (Exemple 2) qui est, en partie, construite à partir de la norme orthographie française ; et les graphies personnelles (Exemple 3) des auteur·rice·s qui sont, souvent, des graphies dites oralisantes.

'Tous les chiens sont heureux.'

1. Totes los gosses son uroses.
2. Toutés lous goussés soun urousés.
3. Toutéÿ louÿ goussés soun urouzés.

Ces variations rendent difficile l'exploitation de tous les textes occitans. Afin de pouvoir les exploiter et en tirer des informations linguistiques, il est nécessaire de construire des outils de traitement automatique qui soient robustes pour traiter ces variations. Nous pensons que les nouvelles générations d'outils automatiques, par exemple les architectures utilisant des réseaux de neurones, se montrent plus robustes sur la variation que les anciennes générations, comme les modèles d'apprentissage supervisé par règles.

## 3 Nouveau corpus de textes occitans pour travailler sur la variation

Ce travail s'inscrit dans cette nécessité d'entraîner des outils pour qu'ils prennent en compte la variation afin qu'ils soient robustes sur l'ensemble des données. Nous avons choisi de nous concentrer sur la construction d'un corpus de textes du département de l'Ariège.

### 3.1 La variation dialectale

Ce département se trouve à la frontière sud de la France avec l'Espagne. C'est une zone de transition linguistique où l'on trouve de nombreuses isoglosses entre deux dialectes occitans,

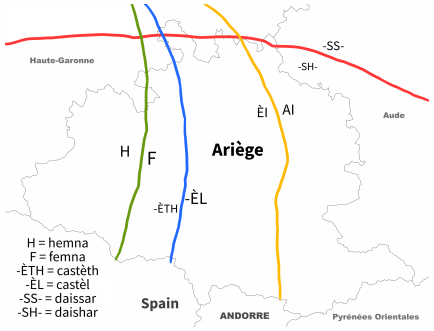


FIGURE 2 – Carte de quelques isoglosses passant par l'Ariège

le gascon (à l'ouest) et le languedocien (à l'est) (Figure 2 [3]). Dans ce département nous trouvons également des parlers "isolés", dans les montagnes, qui ne font pas partie de la zone de transition entre les deux dialectes, mais plutôt entre deux langues : l'occitan et le catalan. Cela fait de ce petit territoire occitan une zone présentant une grande variation dialectale.

Ce n'est pas la seule variation qui nous intéresse dans cette recherche, l'Ariège a aussi la chance de disposer d'une grande production d'écrits occitans. De nombreux·se·s auteur·rice·s ont écrit des contes, légendes, romans ou même des articles de journal en occitan, dans leur parler ariégeois.

## 3.2 La variation graphique

Comme nous l'évoquions plus haut, l'occitan a plusieurs graphies. Dans ce département nous ne pouvons pas quantifier le nombre de graphies existantes, mais nous pouvons les regrouper dans les trois catégories décrites plus haut.

Les locuteur·rice·s de l'occitan n'étant que très rarement alphabétisé·e·s en occitan, nous trouvons de nombreuses personnes ayant l'envie d'écrire dans leur langue mais ne connaissant pas les normes qui peuvent être utilisées. Elles font alors avec les connaissances qu'elles ont, en l'occurrence, en utilisant les graphèmes de la langue dominante, ici, le français. Ces graphies oralisantes sont très intéressantes à étudier afin de mieux connaître la phonologie ou la morphologie de leurs parlers à partir d'écrits.

## 3.3 Constitution du corpus

Afin de construire de tels outils, nous avons constitué un corpus de textes pertinents. Nous avons rassemblé des textes issus des différents groupes de parlers de l'Ariège et des trois

3. Réalisée à partir du fond de carte https ://d-maps.com/carte.php ?num_car=111145&lang=fr

| # tokens | Graphie | | |
|---|---|---|---|
| Dialecte | Classique | Mistralienne | Toutes |
| Languedocien | 10 644 | 7 112 | 17 756 |
| Gascon | 5 710 | 6 749 | 12 459 |
| Autre | 2 936 | 5 163 | 8 099 |
| Tout | 19 290 | 19 024 | 38 314 |

FIGURE 3 – Répartition des tokens dans le corpus Arièja

```
Un      un      un      DET     Gender=Masc|Number=Sing
joun    joun    jorn    NOUN    Gender=Masc|Number=Sing
me      me      me      PRON    Number=Sing|Person=1
passèc  passa   passar  VERB    Number=Sing|Person=3|Mood=Ind|Tense=Past
uno     un      un      DET     Gender=Fem|Number=Sing
ideio   ideio   idèia   NOUN    Gender=Fem|Number=Sing
pel     _       _       _       _       _
per     per     per     ADP     _       _
le      le      lo      DET     Gender=Masc|Number=Sing
cap     cap     cap     NOUN    Gender=Masc|Number=Sing
.       .       .       PUNCT   _       _
```

FIGURE 4 – Annotation d'une phrase du corpus Arièja

types de graphies. Nous avons choisi de regrouper la graphie mistralienne et les graphies oralisantes étant donné qu'elles sont basées sur le rapport graphie-phonie du français.

Notre corpus (Corpus Arièja) est représentatif de toute ces variations théoriquement présente sur le territoire. Certains textes sont assez récents pour en avoir une copie numérique, mais ce n'est pas le cas de tous. Avant de pouvoir les traiter, nous avons dû les numériser dans la quasi entièreté et les océriser.

Le Corpus Arièja est constitué de 38 314 tokens avec 66 textes et 47 auteurs et autrices. Le tableau 3 présente la répartition des tokens dans le corpus en fonction des dialectes et des graphies des textes. Nous avons gardé des proportions similaires pour les deux graphies mais il n'a pas été possible de faire la même chose pour les dialectes.

## 4 Annotation du corpus et outils

Le corpus est annoté en lemmes, parties du discours (POS) et la flexion verbale l'est également. Pour les annotations POS et de la flexion verbale, nous suivons les préconisations de Universal Dependencies (Nivre *et al.*, 2016).

L'annotation des lemmes se fait en deux parties. Un premier lemme qui suit la graphie de l'auteur·rice : nous déduisons du reste du texte le lemme que l'auteur·rice aurait produit. Nous annotons également ce que nous appelons le Supra-lemme, qui est un lemme qui ne suit pas toujours la graphie de l'auteur·rice. Ce Supra-lemme sert simplement à accéder plus facilement aux tokens qui nous intéressent sans que la variation n'interfère. L'illustration 4 montre les différentes annotations présentes dans le corpus.

| Dialecte | % Exactitude |
|---|---|
| Languedocien | 89,49 |
| Limousin | 82,48 |
| Gascon | 81,41 |
| Provençal | 81,37 |
| Tout | 83,32 |

a) Talismane

| Dialecte | % Exactitude |
|---|---|
| Languedocien | 97,30 |
| Provençal | 96,72 |
| Limousin | 94,47 |
| Gascon | 93,65 |
| Tout | 96,75 |

b) AllenÒc

Restaure - Tolosa Treebank

| % Exactitude | Graphie | | |
|---|---|---|---|
| Dialecte | Classique | Mistralienne | Toutes |
| Languedocien | 82,37 | 67,27 | 76,09 |
| Gascon | 76,25 | 67,82 | 72,50 |
| Autre | / | 68,6 | 68,6 |
| Tout | 79,65 | 67,98 | 73,23 |

a) Talismane

| % Exactitude | Graphie | | |
|---|---|---|---|
| Dialecte | Classique | Mistralienne | Toutes |
| Languedocien | 94,44 | 89,79 | 92,74 |
| Gascon | 90,60 | 82,89 | 87,37 |
| Autre | / | 90,71 | 90,71 |
| Tout | 93,26 | 91,32 | 93,35 |

b) AllenÒc

Corpus Arièja

FIGURE 5 – Résultats des modèles sur différents corpus

Le corpus est annoté, pour partie (21 301 tokens) manuellement (à partir d'une pré annotation automatique) et pour partie (17 013 tokens) automatiquement, avec les outils finalisés et ayant de bons résultats sur les variations du corpus.

Nous comparons les résultats de deux modèles d'apprentissage automatique pour déterminer celui qui est le plus robuste face à cette variation. L'un, Talismane (Urieli, 2013; Vergez-Couret & Urieli, 2015) est une architecture qui avait déjà servi pour l'annotation des corpus issus des projets Restaure et Linguatec. C'est un modèle d'apprentissage par règles et probabiliste. L'autre, AllenÒc, utilise des réseaux de neurones et est basé sur la bibliothèque de deep-learning AllenNLP (Gardner *et al.*, 2018).

Ces premiers modèles d'apprentissage automatique, dont nous donnons les résultats ci-après, sont entraînés à partir des corpus Tolosa Treebank (Miletic *et al.*, 2020a) et Restaure (Bernhard *et al.*, 2018) n'intégrant pas de variation graphique, pour faire de l'annotation automatique en parties du discours. Ils seront une nouvelle fois entraînés avec la partie du CorpusArièja corrigée manuellement et qui inclut plusieurs graphies. Ces modèles ont été entraînés à partir de peu de données. Toutefois plusieurs études (Bernhard *et al.*, 2021), ainsi que nos résultats, montrent que nous pouvons obtenir de bons scores d'annotation automatique avec peu de données d'entraînement. Nous montrons que, même avec peu de données d'entraînement, nous pouvons avoir des outils robustes et efficaces pour l'annotation de la plupart des données textuelles.

Les résultats (Figure 5) montrent que le modèle AllenÒc est bien plus performant que le modèle Talismane, que ce soit sur la variation dialectale ou graphique. Les deux modèles ont tendance a être plus performants sur le languedocien et moins sur le gascon. Cela s'explique par le nombre de tokens différents de chacun des dialectes dans les corpus d'entraînement. Le dialecte le plus présent est le languedocien et les autres dialectes sont moins présents. Le test sur le corpus Arièja, avec les différentes graphies, montre que le modèle AllenÒc est plus robuste face aux variations graphiques que Talismane.

Une fois les modèles entraînés sur des corpus contenant de la variation graphique, nous nous attendons à avoir de meilleurs résultats sur la variation graphique. Pour ce qui est d'AllenÒc, nous pensons que les résultats seront bien meilleurs, et rejoindront presque ceux que nous avons sur les textes sans variation graphique. Toutefois, en ce qui concerne Talismane, nous pensons que les résultats ne seront pas beaucoup améliorés. En effet, il est très sensible à la variation au sein de son corpus d'entraînement.

# Références


BEC P. (1995). *La langue occitane*. Que sais-je ? 1059. Paris : Presses universitaires de France, 6e édition corrigée.

BERNHARD D., LIGOZAT A.-L., BRAS M., MARTIN F., VERGEZ-COURET M., ERHART P., SIBILLE J., TODIRASCU A., BOULA DE MAREÜIL P. & HUCK D. (2021). Collecting and annotating corpora for three under-resourced languages of France : Methodological issues. *Language Documentation & Conservation*, **15**, 316–357. HAL : hal-03273196.

BERNHARD D., LIGOZAT A.-L., MARTIN F., BRAS M., MAGISTRY P., VERGEZ-COURET M., STEIBLE L., ERHART P., HATHOUT N., HUCK D., REY C., REYNÉS P., ROSSET S., SIBILLE J. & LAVERGNE T. (2018). Corpora with Part-of-Speech Annotations for Three Regional Languages of France : Alsatian, Occitan and Picard. In *11th edition of the Language Resources and Evaluation Conference*, Miyazaki, Japan. HAL : hal-01704806.

BRAS M., HATHOUT N., SIBILLE J., VERGEZ-COURET M., SÉGUIER A. & DAZEAS B. (2020). Lofl̀oc : Lexic Obert flechit occitan. In J.-F. C. ET DAVID FABIÉ, Éd., *Fidelitats e dissidéncias. Actes del XIIn Congrès de l'Associacion internacionala d'estudis occitans. Actes du XIIe Congrès de l'Association internationales d'études occitanes. Albi 10-15/07/2017*, p. 141–156. Section française de l'Association internationale d'Etudes Occitanes. HAL : hal-03082686.

BRAS M. & VERGEZ-COURET M. (2016). Batel̀oc : A text base for the occitan language.

DIAS G., Éd. (2015). *Actes de TALN 2015 (Traitement automatique des langues naturelles)*, Caen. ATALA, HULTECH.

GARDNER M., GRUS J., NEUMANN M., TAFJORD O., DASIGI P., LIU N., PETERS M., SCHMITZ M. & ZETTLEMOYER L. (2018). Allennlp : A deep semantic natural language processing platform.

MILETIC A., BRAS M., VERGEZ-COURET M., ESHER L., POUJADE C. & SIBILLE J. (2020a). Building a Universal Dependencies Treebank for Occitan. In *12th Language Resources and Evaluation Conference*, p. 2932–2939, Marseille, France. HAL : hal-02892715.

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

NIVRE J., DE MARNEFFE M.-C., GINTER F., GOLDBERG Y., HAJIČ J., MANNING C. D., MCDONALD R., PETROV S., PYYSALO S., SILVEIRA N., TSARFATY R. & ZEMAN D. (2016). Universal Dependencies v1 : A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, p. 1659–1666, Portorož, Slovenia : European Language Resources Association (ELRA).

URIELI A. (2013). *Robust French syntax analysis : reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. Theses, Université Toulouse le Mirail - Toulouse II. HAL : tel-00979681.

VERGEZ-COURET M. & URIELI A. (2015). Analyse morphosyntaxique de l'occitan languedocien : l'amitié entre un petit languedocien et un gros catalan. In *TALARE 2015*, Caen, France. HAL : hal-01214566.