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

Clamença Poujade

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

Clamença Poujade - CLLE (UMR 5263) Joliciel Informatique

Directrice de thèse : Myriam Bras Encadrant CIFRE : Assaf Urieli

Journées du GdR LIFT - 20, 21 novembre 2023

1. Le défi de la variation

La **variation** est un grand enjeu en traitement automatique des langues, d'autant plus pour le traitement automatique des **langues moins ou peu dotées**. En effet, les données les plus facilement traitées sont celles qui sont dans une forme de **standard**. Dès que l'on s'éloigne de ce standard, les outils de traitement automatique ont du mal à effectuer leur traitement. L'**occitan**, langue pouvant être considérée comme moins dotée, connaît beaucoup de variations dans ses textes, par exemple, une **variation dialectale et orthographique**. Dans cette étude, nous nous concentrons sur l'annotation morphosyntaxique, en parties du discours et flexion verbale, et sur la lemmatisation des textes.

PROBLÉMATIQUE : Comment construire des **outils** d'annotation automatique qui soient **robustes** face à des textes contenant de la **variation dialectale et graphique**.
HYPOTHÈSE : Les **nouvelles générations d'architectures**, comme les réseaux neuronaux, d'outils d'annotation automatique sont **plus robustes** sur la variation que les **anciennes générations**, comme des modèles d'apprentissage supervisé par règles.
OBJECTIFS :

- Constituer un **corpus** contenant de la **variation**,
- Observer comment des **outils** d'annotation automatique entraînés sur des données avec **peu de variation** se comportent face à des **données avec variation**,
- Chercher comment **améliorer** les résultats pour avoir des **outils robustes**.

2. L'occitan et ses variations

L'**occitan** est une langue parlée dans le **sud de la France**, dans le **Val d'Aran** dans l'état espagnol et dans certaines **vallées du piémont italien**.

La **variation dialectale** :

- **Six dialectes** (Bec, 1995) (Fig. 1)
- Faisceaux d'**isoglosses** séparant les dialectes
- **Continuum** linguistique

Fig. 1: Carte des dialectes occitans. (J.Sibille pour Bernhard et al., 2021)

Variation graphique:

- Norme dite **classique** (Exemple 1)
- Norme dite **mistralienne** (Exemple 2)
- Graphie **oralisante** (Exemple 3)

'Tous les chats sont heureux.'

1. *Totes los gats son uroses.*
2. *Toutés lous gats soun urousés.*
3. *Touteÿ louÿ gats soun urouzés.*

3. La spécificité de l'Ariège

L'**Ariège** est un département du sud de la France, à la frontière avec l'Andorre et la Catalogne.
Dialectes :

- **Gascon** (ouest)
- **Languedocien** (est)

De nombreux **isoglosses** distingant ces dialectes passent par le département (Fig. 4).

- **Parlers de transition** (gascon, languedocien, catalan)

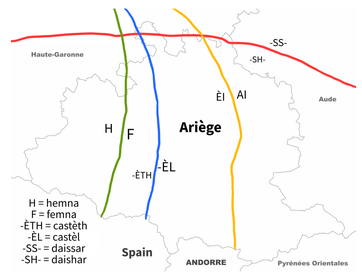

Fig. 3: Quelques isoglosses passant par l'Ariège

Les **variations graphiques** sont similaires à celles de l'occitan plus général, elles s'adaptent aux parlers locaux.

4. Totis les gats son urosis.

L'exemple 4 montre la graphie classique adpatée à un parler, la terminaison de *"totis"* ou *"urosis"* suit la prononciation locale.

4. La constitution du corpus Arièja

- 38 314 tokens
- 47 autrices et auteurs
- 66 textes
- 1850-2022
- Prose

# 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

Fig. 2: Répartition des tokens dans le corpus

5. Les annotations

Universal Dependencies (Nivre et al., 2016) :

- Parties du Discours
- Flexion verbale

Lemmatisation :

- Lemme de l'auteur·rice
- Supralemme

```
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   _       _
```

6. Les annotations POS et résultats des outils automatiques

TALISMANE (Vergez-Couret et Urieli, 2015) : Modèle d'apprentissage par règles et probabiliste

ALLENÒC (adapté de Gardner et al., 2018) : Bibliothèque de deep-learning par réseaux de neurones

Entraînement corpus **Restaure** (Bernhard et al., 2018) et **Tolosa Treebank** (Miletic et al., 2020) **sans variation graphique** mais avec un peu de **variation dialectale.**

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

Fig. 5: Résultats pour les données des corpus Restaure et 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

Fig. 6: Résultats pour les données du corpus Arièja

7. References

Bec, P. (1995). La langue occitane (6e édition corrigée). Presses universitaires de France.
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. https://hal.archives-ouvertes.fr/hal-03273196
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. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), 1659–1666. https://aclanthology.org/L16-1262
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.
Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N. F., Peters, M. E., Schmitz, M., & Zettlemoyer, L. (2018). Allennlp: A deep semantic natural language processing platform. CoRR, abs/1803.07640. http://arxiv.org/abs/1803.07640
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. 11th edition of the Language Resources and Evaluation Conference. https://hal.archives-ouvertes.fr/hal-01704806
Miletic, A., Bras, M., Vergez-Couret, M., Esher, L., Poujade, C., & Sibille, J. (2020). A four-dialect treebank for Occitan: Building process and parsing experiments. Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects, 140–149. https://aclanthology.org/2020.vardial-1.13 La **figure 3** a été réalisée avec le fond de carte https://d-maps.com/carte.php?num_car=111145&lang=fr