More than just data: Dialectal variation and NLP resources for Corsican and Poitevin-Saintongeais

Cristina Garcia Holgado

Plus que des données: La question de la variation dialectale et les ressources en TAL pour le corse et le poitevin-saintongeais

Cristina Garcia Holgado[1,2]
(1) FoReLLIS, UMR 15076, Université de Potiers, 86000 Poitiers, France
(2) LISA, UMR CNRS 6240, Université de Corse - Pasquale Paoli, Corte, France
cristina.garcia.holgado@univ-poitiers.fr

RÉSUMÉ

Le poitevin-saintongeais et le corse ont récemment rejoint la communauté du TALN. Cependant, ces langues sont confrontées à des défis importants en raison de la rareté des corpus annotés, et en outre, elles ne constituent pas des entités uniformes mais présentent des variétés dialectales multiples, malgré les efforts controversés pour établir une référence standardisée. Au cours de la dernière année, elles ont été dotées de ressources linguistiques qui ont conduit aux premières tentatives d'exploration et d'évaluation d'un nombre d'applications TALN. Cependant, mettre en lumière une langue régionale dans le paysage technologique implique également de reconnaître sa réalité linguistique : la diversité dialectale à travers ses différents territoires. Dans ce résumé, nous présentons les premières expériences d'application de méthodes et de techniques supervisées pour gérer le manque d'annotations, en particulier pour le poitevin-saintongeais, et nous soulignons l'importance de couvrir leur variation dialectale, ce que nous cherchons à aborder dans nos travaux futurs.

ABSTRACT

More than just data : Dialectal variation and NLP resources for Corsican and Poitevin-Saintongeais

Poitevin-Saintongeais and Corsican have recently joined the NLP community. However, they face significant challenges due to the scarcity of annotated corpora, and moreover, they do not constitute a homogeneous entity but multiple dialectal varieties, despite the controversial efforts to establish a standardized reference. In the past year, they have been equipped linguistic resources, leading to the first attemps to explore and evaluate a few NLP methods. However, shedding light on a regional language in the technological landscape also entails acknowledging their linguistic reality : the dialectal diversity across it's differents territories. In this summary, we outline initial experiences applying supervised methods and techniques to handle the lack of annotations, specially for Poitevin-Saintongeais, and we highlight the importance of covering their dialectal variation, which we seek to address in future work.

MOTS-CLÉS : corse, poitevin-saintongeais, langues régionales, langues peu dotées, lexiques, variation dialectale.

KEYWORDS: corsican, poitevin-saintongeais, regional languages, low resource language, lexicons, dialectal variation.

1 Context

There is a growing interest in providing digitalised linguistic resources to regional languages in France as shown in (Kevers *et al.*, 2019), (Millour *et al.*, 2017) and (Bernhard *et al.*, 2021). In the case of Corsican and Poitevin-Saintongeais, numerous textual resources are available : The first benefits from an online linguistic database, the BDLC (Banque de Données de la Langue Corse) (Stella Retali-Medori, 2022) which originally included texts from oral sources from different regions, and currently integrates the CCdC (Corpus Canopé de Corse, (Kevers, 2022)) composed of literary and historical texts. Poitevin-Saintongeais benefits from the TELPOS (Dourdet *et al.*, 2019) database, which contains more than 125 bibliographic references for literary texts. The textual resources for Poitevin-Saintongeais are characterized by different spellings, where only a few use the standard spelling. Moreover, a few aligned fragments with other regional languages are available in the ParCoLaF (Miletic *et al.*, 2017) database [1] for both languages. Besides the avaliability of these digital resources in both languages, most of that data remains unannotated and not readily exploitable. In the last year, these two languages have ventured into the NLP domain seeking to develop annotated corpus and lexicons to experiment with supervised approaches.

2 Recent work

2.1 Difficulties in a low resource scenario

Recent approaches in low resource (LR) settings generally rely on a high resource related language, using methods like transfer learning by choosing a suitable transfer language. In this sense, while it may seem intuitive to utilize French for Poitevin-Saintongeais and Italian for Corsican, given their shared Gallo-Romance and Italo-Romance roots respectively, these languages encounter specific challenges to be considered :

1. A **limited availability of linguistic resources**, such as lexicons and dictionaries. Although a few exist, they are subject to copyright whose access is not always guaranteed, or they are based on a particular spelling.
2. **Limited parallel corpus**. Aligned sentences from two literary works are available at ParCoLab, but the amount of aligned data remains very scarce for both languages. However, Poitevin-Saintongeais profits from bilingual articles featured in the journal

1. http ://parcolab.univ-tlse2.fr/corpus/search

Culture Nouvelle Aquitaine [2] since 2023, with uses standardized spelling and provides an opportunity to increase and exploit parallel corpora.

3. Generally, regional languages **lack of standardized spelling** which adds a significant complexity, specially for tasks that are necessary in the area of descriptive linguistics, such as lemmatization. Poitevin-Saintongeais has a normalized spelling, and the Corsican has been recognized as co-official language in Corsica since 2013. However, the texts available in these languages are very diverse, belonging to different geographical areas and therefore, characterised by different variants, speeches (*parlers*) and spellings (*graphie*) that must be taken into account.
4. Along with this, an important challenge arises from the **diverse diffusion areas** (*aires dialectales*) found in regional languages. When we refer to data scarcity, we also encompasses the scarcity of available texts that are annotated with consideration to their various dialectal features that actually constitute those languages. To date, the texts annotated for Poitevin-Saintongeais have been limited to those with standardized spelling so as to address the dialectal variety question in a later stage. For corsican, a set of texts from different sources were annotated regardless of the presence of dialectal variation.
5. Although the BDLC, and particularly the TELPOS database, contain an important number of texts, there is **insufficient metadata** for an effective dialect characterization and processing from an NLP perspective. While a few metadata information on Poitevin-Saintongeais, such as the use of standardized spelling or locality, has been sporadically annotated in some texts, the same attributes are currently unavailable for Corsican texts. Hence, there is still considerable work to annotate these texts in the databases to provide them with linguistic and geographical metadata.

2.2 Approaches

Early work on these languages has focused on overcoming the lack of annotated data. An annotation campaign took place during the last year leading to the first NLP experiences for both languages, which primarily targeted transfer learning methods considering the available resources :

— For Corsican, morpho-syntactic analysis started to be evaluated at LISA (University of Corsica) over an annotated corpus of ~7k tokens. This work, which is in progress, assested the effectiveness of pretraining embeddings in corsican and italian, while using different training sizes. Additional resources, such as parallel corpuses or lexicons, were not used at present, but it could be an envisageable option for the future considering the progressive availability of parallel sentences in ParCoLaF and the constitution of a morphologically inflected lexicon.

2. https ://www.culture-nouvelle-aquitaine.fr/langues-et-cultures-regionales/traduire-le-site-en-langues-regionales/

— Poitevin-saintongeais has followed a similar path including lemmatization. A smaller corpus of ~3k tokens was used, which made this task more arduous. To address this data gap, a lexicon of ~20k entries was compiled via an online bilingual dictionary (Pivetea, 2006), transformed, expanded (~40k entries) and adapted to the respective Universal Dependencies (UD) guidelines. This work had a dual goal : first, to accelerate annotator's decisions by integrating the lexical entries to a collaborative annotation notebook, and second, to generate augmented corpora by transferring new lexical information via distributional neighbours to assess the benefits of a lexicon-based strategy for morpho-syntactic analysis, using both probabilistic (HMM) and neural models (LSTM). This approach has proved to be beneficial without requiring an extensive lexicon. An improvement will be expected with the incorporation of inflected verbs, which were not naturally present in the source dictionary as opposed to nouns and adjectives for which we could provide the inflectional paradigm.

2.3 Limitations

Both lines of work sought to increase the number of annotated texts in order to be able to perform finer NLP tasks, but also to propose the first pos-tagging models for these languages. However, several questions arise at this point : When enlarging the annotated corpus, how well are the distinct linguistic phenomena of these languages represented, and how does the representation of their syntactic and morphological structures impact the quality of predictions ? And most important, how effective are these models when applied to their different dialectal variants ? These questions are intended to show that the quantity of annotated texts is not a sufficient objective in the framework of regional languages, but also the quality.

3 Conclusion and future work

While basic NLP tasks seem straightforward due to the availability of different methods that gradually try to adapt to low resource scenarios, a major challenge arises when addressing dialectal variation. Despite the positive results of this work, none of them have taken yet into account their dialectal dimension. Given the nature of the project they integrate, this stage therefore becomes a required line of research in the work to come. This would enable the representation of their linguistic reality by offering a more nuanced visibility into their different geographical areas. These differences are primarily evident in the lexical and phonetic levels, although they can extend to the morphology and to the idiomatic expressions. In this context, we consider essential to understand that the efforts dedicated to equip these languages go beyond providing enough data for NLP applications. They are motivated by the broader goal of preserving and revitalizing their linguistic heritage. As a result, this undertaking necessitates a comprehensive understanding of their intricate linguistic realities,

and for that, the NLP community requires a strong support from linguistic experts to cover this essential feature. In this sense, the first objective will be to characterize and to identify the variation, and to do so, we will require a significant effort in representing these varieties in the corpus. In short, handling linguistic variation is a central focus of our current work, which has been embedded within a thesis project that will seek to develop tools to ensure their survival and growth, as both Corsican and Poitevin are considered endangered languages by the UNESCO.

4 Acknowledgements

This work was funded by the National Research Agency, via the project DIVITAL (ANR-21-CE27-0004).

Références

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

DOURDET J.-C., VERGEZ-COURET M. & LAY M.-H. (2019). Telpos - Texte électronique en poitevin-saintongeais, enjeux et difficultés. In *Colloque "Langues minoritaires" : quels acteurs pour quel avenir ?*, Strasbourg, France. HAL : hal-02892750.

KEVERS L. (2022). *CCdC - Le Corpus Canopé de Corse*. Rapport interne, UMR 6240 CNRS LISA - Université de Corse. HAL : hal-03912288.

KEVERS L., GUÉNIOT F., GHJACUMINA TOGNOTTI A. & RETALI MEDORI S. (2019). Outiller une langue peu dotée grâce au TALN : l'exemple du corse et BDLC. In E. MORIN, S. ROSSET & P. ZWEIGENBAUM, Éds., *26e Conférence sur le Traitement Automatique des Langues Naturelles*, p. 371–380, Toulouse, France : ATALA. HAL : hal-02567779.

MILETIC A., STOSIC D. & MARJANOVIĆ S. (2017). Parcolab : A parallel corpus for serbian, french and english. In *International Conference on Text, Speech and Dialogue*.

MILLOUR A., FORT K., BERNHARD D. & STEIBLÉ L. (2017). Vers une solution légère de production de données pour le TAL : création d'un tagger de l'alsacien par crowdsourcing bénévole. In *Traitement Automatique des Langues Naturelles (TALN)*, Orléans, France. HAL : hal-01516226.

PIVETEA V. (2006). *Dictionnaire francais poitevin-saintongeais*. Geste Éditions.

STELLA RETALI-MEDORI L. K. (2022). La morphologie dans la banque de données langue corse : bilan et perspectives. *OpenEdition*.