Beyond Standardization: Crafting a Comprehensive Annotated Corpus for Corsican and Poitevin-Saintongeais

Cristina García Holgado, PhD student
University of Corsica Pasquale Paoli (France)
University of Poitiers (France)

Keywords: *Corsican, Poitevin-Saintongeais, Low-resource languages, Corpora, Variation*

Corsican and Poitevin-Saintongeais are two languages of France that have recently become part of the NLP community. The first is a continuum of four to five dialectal variants spoken mainly on the island of Corsica, and fits naturally into the Italo-Romance family. Its spelling is not standardized by a consensual norm[1] and embraces the concept of polynomic language (*langue polynomique*). The second is a Gallo-Romance language spoken between the Loire and Garonne rivers, which exhibits a strong Occitan influence while sharing morphological similarities with French, thus distinguishing two major dialectal areas: Poitevin and Saintongeais. A standardized spelling norm was proposed, but its use sparks much debate as it is seen as contrived by the speakers. In essence, neither language constitutes a single entity but rather a blend of dialects and local varieties, varying in geographical granularity. This must be taken into account during the process of providing these languages with NLP resources, a key aspect being investigated in an ongoing thesis project.

Recent work for both languages has led to the creation of an small annotated corpora in CoNLL-U format and the creation of an exploitable lexicon compiling some dialectal traits (only Corsican for the moment). Corsican disposes of 220 annotated sentences while Poitevin-Saintongeais 130 and both have been equipped with a lexicon, at present requiring a few enhancements. It must be noted that the project's objective is not solely to equip these languages for the development of language technologies but primarily to preserve their linguistic heritage. This emphasizes the necessity of considering their complete dialectal framework.

While there are UD guidelines developed for POS tagging, some linguistic questions need consideration in the annotated corpora. As a result, our project concerns particularly the enhancement of an annotated corpus which is intended to lead to the development of a Treebank. These enhancements aim to address the following aspects:

- First, exploring the possibility of managing specific agglomerated and contracted forms within the annotations: According to the UD guidelines, the sentences should be split into “syntactic” words, including the separation of clitics and decomposition of contractions.

For instance, as mentioned in[1] for Corsican, some pronominal verbs exist in their agglomerated form (eg. spassassi ‘to have fun’) and separated by a space (spassà si), or the irregular use of punctuation signs within tokens (eg. cum’è, which is the single token ‘like’). In Poitevin-Saintongeais, the contraction of words is common in some variants as “*p’tit*” (“petit” = “small”).

- A second element for this project concerns the link of the given lemma to their possible realizations: There is no consensus on the canonical base form (lemma) given for a form. For example, the chestnut blossom is mainly called *trama* or *tràmula* in Corsica. Occasionally, other forms are used, such as *lianda*, *calchera, misgi misgi*, and *chjuchjurulanda*[2]. While the French word *quelques* (“some”) has the lemmas *çhauques, quauques, çheùques, queùques, çhéques,* and *quéques* in Poitevin-Saintongeais. In this regard, the need to choose a supralema is being considered for processing reasons, possibly accompanied by an additional annotation layer that captures (for example, a layer for morphological varieties of lemmas) or links to the different variants of the lemmas (with an external lexicon) in order to converge the computational and linguistic approaches.

- Thirdly, to ponder an additional dialectal layer that incorporates some of the regional or local specificities: This is particularly important for Corsican as a portion of the corpus consists of ethnotexts[3] (testimonies from speakers about regional customs). These texts contain idioms and specialized vocabulary unique to certain domains, showcasing distinctive morphological variations influenced by their dialectal areas (e.g.: *castanghju* in the north/*pulloni* in the south):

  — Un *castanghju* giovanu, cumu si chjama ?
  — [...] A ghjente punìa i *pulloni* è i *pulloni*

In summary, this project aims to explore ways to capture and document dialectal diversity within the UD annotated corpora and develop a new Treebank for these languages. For this reason, joining this event would be practical to clear up uncertainties about the annotations, discuss the project with fellow researchers or participants tackling similar issues, and potentially discover new leads, especially regarding the dialectal scope which is still at a very early age of development.

References

[1] Alice Millour, Laurent Kevers, Lorenza Brasile, Alberto Ghia. Agettivu, aggitivu o aghjettivu? POS Tagging Corsican Dialects. *LREC-COLING 2024*, May 2024, Turin, Italy. ⟨hal-04534608⟩

[2] Marie-José Dalbera Stefanaggi, Stella Retali Medori. Castagni è puddoni, la castanéiculture en Corse : lexique et usages. Editions Sammarcelli, 206 p., 2013. ⟨hal-00990380⟩

[3] Stella Retali Medori, Laurent Kevers. La morphologie dans la Banque de Données Langue Corse : bilan et perspectives. Corpus, 2022, Corpus et données en morphologie, 23, ⟨10.4000/corpus.7115⟩. ⟨hal-03591866⟩