



**HAL**  
open science

## Tooling up a less-resourced language with NLP : the example of Corsican and the BDLC

Laurent Kevers, Stella Retali Medori, Florian Gueniot, A. Ghjacumina Tognotti

► **To cite this version:**

Laurent Kevers, Stella Retali Medori, Florian Gueniot, A. Ghjacumina Tognotti. Tooling up a less-resourced language with NLP : the example of Corsican and the BDLC. International Conference on Language Technologies for All (LT4All) Enabling Linguistic Diversity and Multilingualism Worldwide, Dec 2019, Paris, France. hal-02880729

**HAL Id: hal-02880729**

**<https://hal.science/hal-02880729>**

Submitted on 25 Jun 2020

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Roadmap

- Corpora & consultation interface (concordancer)
- Electronic lexical resource for NLP
- Language detection tool
- Part-of-speech (POS) tagger

## Tooling up a less-resourced language with NLP : the example of Corsican and the BDLC

Kevers L., Retali-Medori S., Guéniot F., Tognotti A.G.  
 UMR CNRS 6240 LISA - Università di Corsica



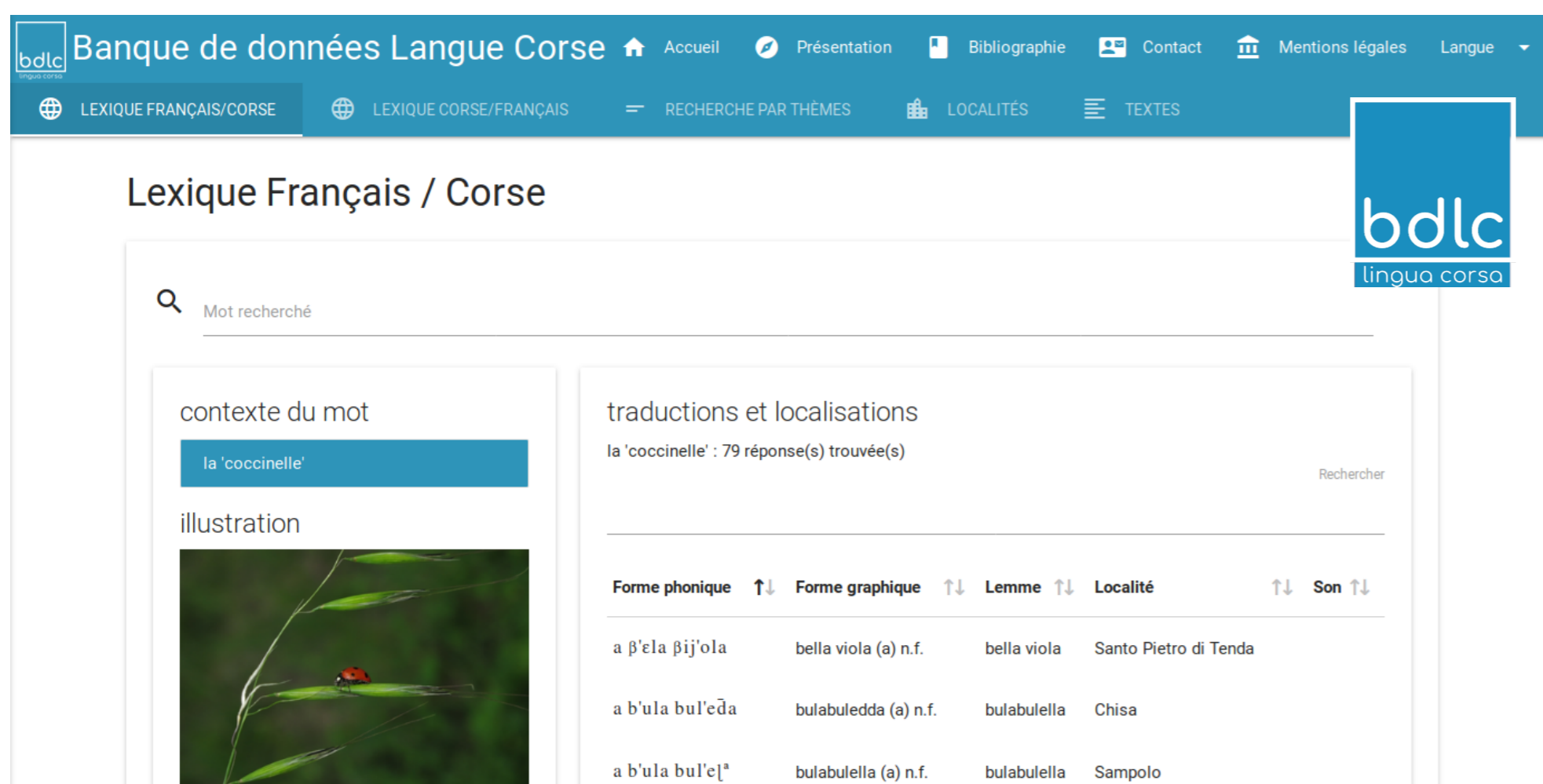
### 1. The Corsican language & BDLC

- Latin language
- Italo-Romance domain
- Borrowings from various languages
- 4 to 5 dialectal areas
- Unstandardized writing
- Diglossia with French

#### BDLC (corsican language database): Linguistic data related to Corsican know-how and cultural traditions

- Field surveys
- Thematic questionnaires (FR wordlists)
  - > Corsican translations
  - > Ethnotexts (testimonies)
- Important variation
  - \* an object can be related to many lemmas
  - \* variations in transcripts :
    - keep the dialectal richness : *cerra, gerra*
    - hiatus accentuation : *durmìa*
    - enclitics : *fanne = fà* («to do») + *ne* («it»)
    - opening of the proparoxytones vowels : *pèrgula*

- URL: <http://bdlc.univ-corse.fr/>



### 2. Corpora collection

- More than 3MIO words collected from various online sources (Wikipedia, The Bible, press, blogs...)
- Usefull for the research usage of our team
- Most of the copora cannot be redistributed because of legal (licensing, copyright) issues
- Attempt to gather exploitable and redistributable corpora for all (eg. under license CC BY-NC-SA 4.0)

### 3. Corpora interface

- Online concordancer, linguistic search criteria (WIP)
- In the future : lemmatized corpora

### 4. Electronic dictionary

- DELA format: `form,lemma.gram_sem_codes:flex_codes/comment`
- Content : 20 875 forms
  - 17 860 simple forms (10 224 lemmas)
  - 3 015 compound forms (2 244 lemmas)
- Coverage : 49% (corpus: 160K forms / 15K unique)
- To be extended...
- Pending : how to deal with dialectal variation ?
  - > Lemma choice, use of upper level lemma... ?

### 5. Lemmatization procedure

- Based on the dictionary
- Two steps :
  1. Lexicographic study
  2. Disambiguation
- To be implemented under Unitex

### 6. Language detection tool

- No ready-to-use (good) software including Corsican
- Re-training and testing of different systems

8 languages :	Method	8 languages	Corsican	9 languages
- English	MyLetterDistrib	99.62	93.04	98.89
- German	MyStopWords	99.62	93.56	98.95
- Dutch	CueLanguage	99.50	84.41	97.82
- French	Libre TextCat	100	95.62	99.51
- Italian	Langid.py	98.75	95.23	98.36
- Spanish	Langdetect	100	96.65	99.63
- Portuguese	Langdetect	100	96.65	99.63
- Romanian	FastText	100	95.49	99.50
+ Corsican !	Lgid	100	98.58	99.84

- DEMO: <http://bdlc.univ-corse.fr/ta1/>

### 7. Part-of-speech tagging

Possible options :

- Transfer methods (from Italian)
- Training of a Corsican version (long term)