1. The Corsican Language & BDLC

- Latin language
- Italo-Romance domain
- Borrowings from various languages
- 4 to 5 dialectal areas
- Unstandardized writing
- Diglossia with French
- 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») and ne («it»)
    - Opening of the proparoxytones vowels: pèrgula

BDLC (Corsican Language Database): Linguistic data related to Corsican know-how and cultural traditions

- 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») and 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...)
- Useful for the research usage of our team
- Most of the corpora cannot be redistributed because of legal (licensing, copyright) issues
- Attempt to gather exploitable and redistributable corpora for all (e.g. 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:
  - 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

<table>
<thead>
<tr>
<th>8 languages</th>
<th>Corsican</th>
<th>9 languages</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>99.62</td>
<td>93.04</td>
</tr>
<tr>
<td>MyStopWords</td>
<td>99.62</td>
<td>93.56</td>
</tr>
<tr>
<td>CnLanguage</td>
<td>99.81</td>
<td>95.34</td>
</tr>
<tr>
<td>LibreTextCat</td>
<td>100.00</td>
<td>95.62</td>
</tr>
<tr>
<td>Langdetect</td>
<td>98.75</td>
<td>95.23</td>
</tr>
<tr>
<td>FastText</td>
<td>100.00</td>
<td>96.65</td>
</tr>
<tr>
<td>LgId</td>
<td>100.00</td>
<td>95.49</td>
</tr>
</tbody>
</table>

+ Corsican!

- DEMO: http://bdlc.univ-corse.fr/tal/

7. Part-of-Speech Tagging

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