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

Poster TALN-RECITAL

Outiller une langue peu dotée grâce au TALN : l'exemple du corse et de la BDLC

Feuille de Route

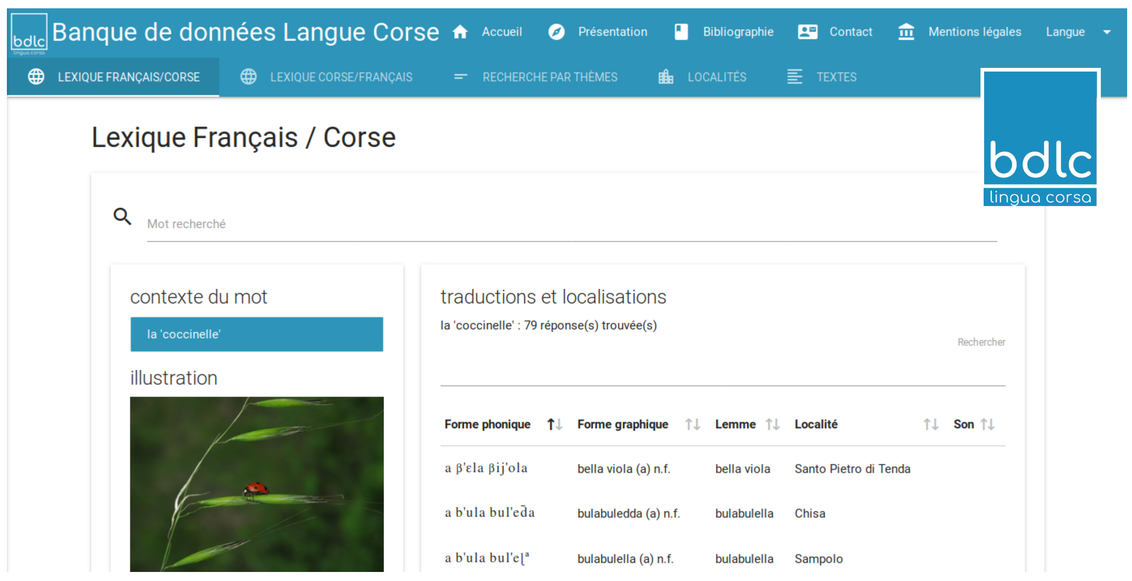
- ☐ **Dictionnaire électronique pour le TAL**
- ☐ **Interface de consultation de textes lemmatisés**
- ☐ **Outil de détection de langue**
- ☐ **Outil d'annotation morpho-syntaxique (POS)**

1. La langue corse et la BDLC

- Langue issue du latin
- Domaine italo-roman
- Emprunts à diverses langues
- 4 à 5 aires dialectales
- Écriture non normée
- Diglossie avec le français

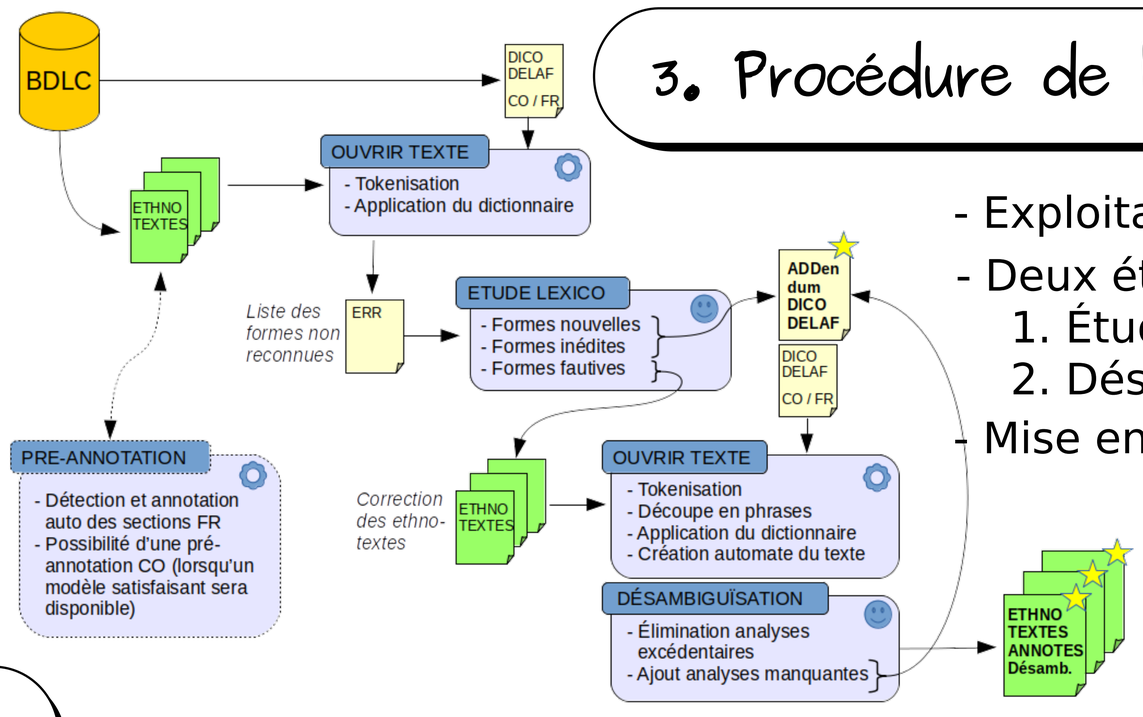
BDLC : Données linguistiques en lien avec les savoir-faires et traditions culturelles corses

- Enquêtes de terrain
- Questionnaires thématiques (listes mots FR)
  -> traductions en corse
  -> ethnotextes (témoignages)
- Variation riche
  * un signifié peut être relié à de nombreux lemmes
  * transcriptions variables :
    - valoriser la variation : *cerra* et *gerra*
    - accentuation des hiatus : *durmìa*
    - enclitiques : *fanne* = *fà* («faire») + *ne* («en»)
    - aperture des voyelles des proparoxytons : *pèrgula*
- Site web : `http://bdlc.univ-corse.fr/`



2. Dictionnaire électronique

- Format DELA :
  `forme,lemme.codes_gram_sem:code_flex/commentaire.`
- 20.875 formes : 17.860 formes simples (10.224 lemmes)
  3.015 formes composées (2.244 lemmes)
- Couverture : environ 49 % des occurrences d'un corpus de ~160.000 formes, dont un peu moins de 15.000 uniques



3. Procédure de lemmatisation

- Exploitation du dictionnaire
- Deux étapes :
  1. Étude lexicographique
  2. Désambiguïsation
- Mise en oeuvre sous Unitex

4. Constitution de Corpus

Nom	Nb. mots	Nb. docs	Poids (Ko)
Déclaration des droits de l'homme	1.977	1	14
Bible	146.489	1	833
Tarrori E Fantasia (nouvelles)	111.310	111	670
Corpus BDLC (01/2019)	83.654	787	458
Canopé de Corse	453.072	37	3.482
M3C Parcours	74.123	107	424
Bonanova (revue littéraire)	531.232	27	2.982
A Piazzetta (journal en ligne)	796.991	1.139	4.742
Wikipedia Corse	926.674	5.692	5.618
Divers	35.514	2	208
TOTAL	3.161.036	7.904	19.431

5. Détection de la langue

- Pas de logiciel satisfaisant incluant le corse
- Ré-entrainement de différents systèmes

Méthode	8 langues	Corse	9 langues
MyLetterDistrib	99,62	93,04	98,89
MyStopWords	99,62	93,56	98,95
CueLanguage	99,50	84,41	97,82
LibreTextCat	100	95,62	99,51
Langid.py	98,75	95,23	98,36
Langdetect	100	96,65	99,63
FastText	100	95,49	99,50
Ldig	**100**	**98,58**	**99,84**

8 langues :
- Anglais
- Allemand
- Néerlandais
- Français
- Italien
- Espagnol
- Portuguais
- Roumain

+ Corse !

6. Annotation morphosyntaxique

- Test en cours avec TreeTagger italien
- Entraînement à terme d'une version corse

7. Interface de consultation

Concordancier, critères de recherche linguistiques