A Survey of Language Technologies Resources and Tools for Corsican

Research report
May 2021

Laurent Kevers (kevers_l@univ-corse.fr)

With the contribution of :

Stella Retali-Medori
Ghjacumina A. Tognotti

UMR CNRS 6240 LISA
Università di Corsica

Introduction

The aim of this survey is to list and outline the resources and tools that can be used in the context of language technologies. The resources and tools that are included in this inventory have the common feature of having the potential to be integrated or to be involved in the development of tools based on these technologies. However, the possibilities for practical exploitation may be limited for several reasons : availability and access to the resource, non-standard formats or legal and licensing issues. We have not included in this study resources that were too far from an operationalisation stage. For example, many sites that offer more or less advanced grammar or vocabulary concepts, and which constitute useful sources of information about the language, have not been included.

The data from this survey has been provided and incorporated into the European project named ELE (European Language Equality, https://european-language-equality.eu) thanks to European Language Equality Network (ELEN, www.elen.ngo).

Sources

The information collected for this document has been compiled by the authors. However, several studies or Internet portals were consulted in order to maximise and complete the data.

- ELDA/DGLFLF Survey (2014) : https://www.culture.gouv.fr/Sites-thematiques/Langue-francaise-et-langues-de-France/Politiques-de-la-langue/Langues-et-numerique/Les-technologies-de-la-langue-et-la-normalisation/Inventaire-des-ressources-linguistiques-des-langues-de-France;
- OLAC: Open Language Archives Community : www.language-archives.org, or the Language Archive Service : https://language-archives.services/olacvis/;
- Clarin, Virtual Language Observatory : https://vlo.clarin.eu;
- European Language Grid : https://live.european-language-grid.eu/;
- Meta Share : http://www.meta-share.org/;
- ELRA Catalogue : http://catalogue.elra.info;
- LRE Map (no Corsican data found) : https://lremap.elra.info;
- LDC (no Corsican data found) : https://www.ldc.upenn.edu/;
- Ortolang (no Corsican data found) : https://www.ortolang.fr;
- Omniglot : https://omniglot.com;
- Lexilogos : https://www.lexilogos.com.

This study refers to data concerning Corsican. We have also come across some resources which are not currently concerned with this language, but which could include it in the future. In particular, there are several large multilingual corpora without Corsican data which could be checked for future versions (if any).

- CC-100 : http://data.statmt.org/cc-100/
- CCAligned : https://opus.nlpl.eu/CCAligned.php
- MultiCCAligned : https://opus.nlpl.eu/MultiCCAligned.php
- OSCAR : https://oscar-corpus.com/
- ParaCrawl : https://paracrawl.eu/; https://opus.nlpl.eu/ParaCrawl.php
- MultiParaCrawl : https://opus.nlpl.eu/MultiParaCrawl.php
- WikiMatrix : https://github.com/facebookresearch/LASER/tree/master/tasks/WikiMatrix

For Technology readiness levels (TRL), see
https://ec.europa.eu/research/participants/data/ref/h2020/wp/2014_2015/annexes/h2020-wp1415-annex-g-trl_en.pdf

Updates

Any corrections, clarifications or additions to the data contained in this survey can be communicated to the authors (kevers_l@univ-corse.fr).

List of tools and resources

1. Lexical and Conceptual Resources

1.1. Banque de Données Langue Corse

Resource name	Banque de Données Langue Corse
Resource type	lexical/conceptual resource
Ressource short name	BDLC
Landing page	https://bdlc.univ-corse.fr
Description	The *Banque de Données Langue Corse* (Corsican Language Data Bank or BDLC) is a scientific tool designed to illustrate and study the language variation of Corsican in space - including the Corsican dialect of Gallura and the alloglot community of Bunifaziu (Ligurian) - and in time. The data contained in the database are the result of field surveys and were collected from native speakers. It includes mainly lexical data, wich could be displayed on maps, but also ethnotexts, audio recordings and pictures.
(Funding) project	Banque de Données Langue Corse (BDLC), Université de Corse, CNRS
Language(s)	Corsican
Resource publication year	1986 - 2021
Lexical / Conceptual Resource Subclass	phonological lexicon; lexicon; terminological resource
Encoding level	morphology; phonetics; phonology; etymology
Media type(s) of parts	text; audio; images

1.2. INFCOR, Banca di dati di a lingua corsa

Resource name	**INFCOR, Banca di dati di a lingua corsa**
Resource type	lexical/conceptual resource
Ressource short name	INFCOR
Landing page	http://infcor.adecec.net/
Description	Portal allowing the interrogation of a lexical database including all the components of the language: traditional vocabulary in its varieties, specific ancient and modern terminologies, proper nouns, locutions... Each entry includes, in addition to the definition in Corsican, the figurative pronunciation, the etymology, the synonyms, the antonyms, the derivatives and compounds, the analogies, as well as the French, Italian and English equivalents. An illustration taken from literary works and, if necessary, a bibliography, complete each article. This resource is only available online (no download) and is copyrighted.
(Funding) project	ADECEC
Language(s)	Corsican (+ French, Italian and English translations)
Resource publication year	2013 (v.5)
Lexical / Conceptual Resource Subclass	dictionary
Encoding level	morphology; phonetics
Media type(s) of parts	texts

1.3. ADECEC's lexicons

Resource name	ADECEC’s lexicons
Resource type	lexical/conceptual resource
Ressource short name	n/a
Landing page	http://adecec.net/html/download.html
Description	Collection of specialised lexicons in various fields (mainly French-Corsican). Some lexicons are available with English or German translations. • Die korsische Sprache (2006) : Corsican-German • U franghju – Le pressage de l'huile (2002) : Corsican-French • A Caccia in Corsica – La chasse (1987) : Corsican-French • A Frutta – Les fruits (1984) : Corsican-French • La Cuisine (1988) : Corsican-French • La Géographie (1983) : French-Corsican ; Géographie 2 (1983) : French-Corsican-English • Le Football (1983) : French-Corsican ; Le Football 2 (1983) : French-Corsican-English • La Presse Audiovisuelle (1983) : French-Corsican • Attenti Trappule... ! (1986) : Corsican-French • Da a Lingua Francese à a Lingua Corsa (1995) : French-Corsican • La Chaudronnerie (1984) : French-Corsican ; La Chaudronnerie 2 (date inconnue) : French-Corsican-English • L'automobile (1978) : French-Corsican • Les Oiseaux (1982) : French-Corsican ; Les oiseaux 2 (1982) : French-Corsican-English • La Philosophie (1981) : French-Corsican ; La Philosophie 2 (1981) : French-Corsican-English • La Linguistique (1981) : French-Corsican • La Maison (1984) : French-Corsican-English • A Vigna è u Vinu (1993) : French-Corsican • Le Temps qu'il fait (1985) : French-Corsican • L'électronique (1980) : French-Corsican • La Psychanalyse (1981) : French-Corsican • Le Droit (1983) : French-Corsican • La Mathématique (1979) : French-Corsican • Ghjochi è Ghjoculi (1989) : Corsican-French

	• Filage et Tissage (1980) : French-Corsican ; Filage et tissage 2 (1980) : French-Corsican-English • Amore è Sessualità (1991) : Corsican-French • Lessicu Martinu Sampieri (date inconnue) : Corsican-French
(Funding) project	ADECEC
Language(s)	Corsican (+French, +English, +German)
Resource publication year	1978 - 2006 (see description)
Lexical / Conceptual Resource Subclass	terminological resource; lexicon
Encoding level	n/a
Media type(s) of parts	texts

1.4. A lingua corsa's lexicons

Resource name	A lingua corsa lexicons
Resource type	lexical/conceptual resource
Ressource short name	n/a
Landing page	https://gbatti-alinguacorsa.pagesperso-orange.fr/lexiques/lexiques.htm
Description	Set of general and specialised lexicons.
(Funding) project	n/a
Language(s)	Corsican
Resource publication year	2018
Lexical / Conceptual Resource Subclass	terminological resource; lexicon
Encoding level	n/a
Media type(s) of parts	texts

1.5. JULIELab/MEmoLon

Resource name	**JULIELab/MEmoLon**
Resource type	lexical/conceptual resource
Ressource short name	n/a
Landing page	http://doi.org/10.5281/zenodo.3756607
Description	This dataset contains the resulting lexicons from a ACL 2020 paper "Learning and Evaluating Emotion Lexicons for 91 Languages". It constitutes a large-scale emotion lexicon, covering roughly between 100k and 2M word type entries, depending of the language. Each word is described in terms of eight emotional variables: valence, arousal, dominance, joy anger, sadness, fear, and disgust. Available under CC Attribution 4.0 International licence.
(Funding) project	n/a
Language(s)	Corsican (amongst 91 languages)
Resource publication year	2020
Lexical / Conceptual Resource Subclass	computational lexicon
Encoding level	semantics
Media type(s) of parts	texts

1.6. BabelNet

Resource name	BabelNet
Resource type	lexical/conceptual resource
Ressource short name	n/a
Landing page	https://babelnet.org
Description	BabelNet is a multilingual encyclopedic dictionary, with wide lexicographic and encyclopedic coverage of terms, and a semantic network/ontology which connects concepts and named entities in a very large network of semantic relations, made up of about 20 million entries. BabelNet 5.0 covers 500 languages, including Corsican, and is obtained from the automatic integration of several resources such as WordNet, Wikipedia, Wiktionary, Wikidata, GeoNames…
(Funding) project	Sapienza University of Rome, Babelscape
Language(s)	Corsican (amongst 500 languages)
Resource publication year	2021 (v.5)
Lexical / Conceptual Resource Subclass	Multilingual semantic network : computational lexicon; machine readable dictionary; mapping of resources; wordnet
Encoding level	phonetics; semantics
Media type(s) of parts	texts; images

1.7. Wikizziunariu, the Corsican Wiktionary

Resource name	Wikizziunariu, the Corsican Wiktionary
Resource type	lexical/conceptual resource
Ressource short name	Wikizziunariu
Landing page	https://co.wiktionary.org/wiki/Pagina_maestra
Description	A collaborative project to produce a free-content multilingual dictionary in Corsican. Content released under the CC BY-SA 3.0 licence.
(Funding) project	Wikimedia
Language(s)	Corsican
Resource publication year	2021
Lexical / Conceptual Resource Subclass	dictionary
Encoding level	morphology; phonetics
Media type(s) of parts	texts; images

1.8. Educorsica Lessicu

Resource name	Educorsica Lessicu
Resource type	lexical/conceptual resource
Ressource short name	Educorsica
Landing page	https://www.educorsica.fr/lessicu/index.php
Description	This lexical repository provides teachers, students and, more broadly, the general public with a corpus of specialised words in several subject areas (history, geography, computing, mathematics, chemistry, art, etc.).
(Funding) project	Canopé, Collectivité de Corse
Language(s)	Corsican; French
Resource publication year	2021 (last update)
Lexical / Conceptual Resource Subclass	terminological resource; lexicon
Encoding level	n/a
Media type(s) of parts	texts

1.9. Parlami Corsu Traduttori

Resource name	**Parlami Corsu Traduttori**
Resource type	lexical/conceptual resource
Ressource short name	Parlami Corsu
Landing page	https://www.parlamicorsu.com/
Description	Online translation dictionary (French-Corsican, Corsican-French)
(Funding) project	Sirviziu Lingua è Cultura Corsa di a Cità d’Aiacciu
Language(s)	Corsican; French
Resource publication year	2014
Lexical / Conceptual Resource Subclass	lexicon
Encoding level	n/a
Media type(s) of parts	texts

1.10. Glosbe, French-Corsican dictionary

Resource name	Glosbe, French-Corsican dictionary
Resource type	lexical/conceptual resource
Ressource short name	Glosbe
Landing page	https://fr.glosbe.com/fr/co
Description	Online translation dictionary (French-Corsican, Corsican-French)
(Funding) project	n/a
Language(s)	Corsican; French
Resource publication year	2021
Lexical / Conceptual Resource Subclass	lexicon
Encoding level	n/a
Media type(s) of parts	texts

1.11. Freelang – Corsican-French-Corsican online dictionary

Resource name	**FREELANG – Corsican-French-Corsican online dictionary**
Resource type	lexical/conceptual resource
Ressource short name	Freelang
Landing page	https://www.freelang.com/enligne/corse.php
Description	Online translation dictionary (French-Corsican, Corsican-French) Also available as a desktop version (Windows)
(Funding) project	n/a
Language(s)	Corsican; French
Resource publication year	1999-2017
Lexical / Conceptual Resource Subclass	lexicon
Encoding level	n/a
Media type(s) of parts	texts

1.12. Parlami Corsu Cunghjucatori

Resource name	Parlami Corsu Cunghjucatori
Resource type	lexical/conceptual resource
Ressource short name	n/a
Landing page	https://www.parlamicorsu.com/
Description	Online service to conjugate a verb.
(Funding) project	Sirviziu Lingua è Cultura Corsa di a Cità d'Aiacciu
Language(s)	Corsican
Resource publication year	2014
Lexical / Conceptual Resource Subclass	lexicon
Encoding level	morphology
Media type(s) of input	text

1.13. Cunghjugatori corsu

Resource name	**Cunghjugatori corsu**
Resource type	tool / service
Ressource short name	n/a
Landing page	https://aiaccinu.eu.org/cunghjugatori/
Description	Web application for the conjugation of Corsican verbs.
(Funding) project	n/a
Language(s)	Corsican
Resource publication year	?
Lexical / Conceptual Resource Subclass	lexicon
Encoding level	morphology
Media type(s) of input	text

2. Written Corpora

2.1. A Piazzetta, giurnale in lingua corsa : a XML TEI Corpus

Resource name	**A Piazzetta, giurnale in lingua corsa : a XML TEI Corpus**
Resource type	corpus
Ressource short name	A Piazzetta
Landing page	https://bdlc.univ-corse.fr/tal/index.php?page=res
Description	Corpus of texts from the journalistic blog *A Piazzetta*. Articles published between December 2010 and September 2019. Size: >504K tokens; Format: XML TEI P5; License: CC BY-NC-SA 4.0
(Funding) project	Banque de Données Langue Corse (BDLC), Université de Corse, CNRS
Language(s)	Corsican
Resource publication year	2019
Corpus subclass	raw corpora
Media type(s) of parts	texts

2.2. A Sacra Bìbbia : a XML TEI Corpus

Resource name	A Sacra Bìbbia : a XML TEI Corpus
Resource type	corpus
Ressource short name	A Sacra Bìbbia
Landing page	https://bdlc.univ-corse.fr/tal/index.php?page=res
Description	Includes all 66 books of the Old and New Testaments. The divisions into books, chapters and verses have been preserved. Size: >771K tokens; Format: XML TEI P5; License: CC BY-NC-SA 4.0
(Funding) project	Banque de Données Langue Corse (BDLC), Université de Corse, CNRS
Language(s)	Corsican
Resource publication year	2019
Corpus subclass	raw corpora
Media type(s) of parts	texts

2.3. Wikipedia, enciclopedia libara in lingua corsa : a XML TEI Corpus

Resource name	**Wikipedia, enciclopedia libara in lingua corsa : a XML TEI Corpus**
Resource type	corpus
Ressource short name	Corsican Wikipedia
Landing page	https://bdlc.univ-corse.fr/tal/index.php?page=res
Description	Corpus containing the Corsican version of the collaborative online encyclopedia Wikipedia (https://co.wikipedia.org). Only the content pages have been kept, and these have been cleaned of "wiki" codes. The corpus is based on the wikipedia dump of 20-10-2019. Size: >919K tokens; Format: XML TEI P5; License: CC BY-SA 3.0 Note: another (multilingual) version of the Wikipedia corpus is also available from http://hdl.handle.net/11234/1-2735 : Wikipedia plain text data obtained from Wikipedia dumps with WikiExtractor in February 2018 (297 languages including Corsican).
Funding) project	Banque de Données Langue Corse (BDLC), Université de Corse, CNRS
Language(s)	Corsican
Resource publication year	2019
Corpus subclass	raw corpora
Media type(s) of parts	texts

2.4. Dichjarazioni Univirsali di I Diritti di L’Omu

Resource name	**Dichjarazioni Univirsali di I Diritti di L’Omu**
Resource type	corpus
Ressource short name	Corsican UDHR
Landing page	https://www.ohchr.org/EN/UDHR/Pages/Language.aspx?LangID=coi or https://www.unicode.org/udhr/translations.html?search=Corsican
Description	The Universal Declaration of Human Rights (UDHR). The particularity of this text is it’s availability in a large number of languages. Unicode.org offers several formats (TXT, XML…). This work is excerpted from an official document of the United Nations. The policy of this organisation is to keep most of its documents in the public domain in order to disseminate "as widely as possible the ideas (contained) in the United Nations Publications". Pursuant to UN Administrative Instruction ST/AI/189/Add.9/Rev.2 available in English only, this document are in the public domain worldwide.
(Funding) project	n/a
Language(s)	Corsican
Resource publication year	1998 (OHCR) / 2006 (Unicode.org)
Corpus subclass	raw corpora
Media type(s) of parts	texts

2.5. Wikipedia in lingua corsa

Resource name	Wikipedia in lingua corsa
Resource type	corpus
Ressource short name	n/a
Landing page	https://co.wikipedia.org/wiki/Pagina_maestra
Description	A free, multilingual online encyclopedia written and maintained by a community of volunteer contributors through a model of open collaboration, using a wiki-based editing system. Available as XML dumps (https://dumps.wikimedia.org/cowiki). Data released under the CC BY-SA 3.0 licence.
(Funding) project	Wikimedia
Language(s)	Corsican
Resource publication year	2021
Corpus subclass	raw corpora
Media type(s) of parts	texts, images

2.6. An Crúbadán – Corpus Building for Minority Languages – Corsican

Resource name	**An Crúbadán – Corpus Building for Minority Languages – Corsican**
Resource type	corpus
Ressource short name	An Crúbadán
Landing page	http://crubadan.org/languages/co
Description	The aim of the Crúbadán project is the creation of text corpora for a large number of under-resourced languages by crawling the web. The Corsican part of this project contains 806 103 words from 150 different documents. The data is available as character trigrams, word bigrams or single words. Data under the CC BY 4.0 licence.
(Funding) project	n/a
Language(s)	Corsican
Resource publication year	2018
Corpus subclass	raw corpora
Media type(s) of parts	texts

2.7. W2C – Web to Corpus

Resource name	**W2C – Web to Corpus**
Resource type	corpus
Ressource short name	W2C
Landing page	https://lindat.mff.cuni.cz/repository/xmlui/handle/11858/00-097C-0000-0022-6133-9 http://hdl.handle.net/11858/00-097C-0000-0022-6133-9
Description	A set of corpora for 120 languages automatically collected from wikipedia and the web. The corsican part of W2C contains 2 767 495 tokens (evaluation made with wc command under Linux). Data available under the CC BY-SA 3.0 licence.
(Funding) project	n/a
Language(s)	Corsican (amongst 120 languages)
Resource publication year	2011
Corpus subclass	raw corpora
Media type(s) of parts	texts

2.8. Common Crawl

Resource name	Common Crawl
Resource type	corpus
Ressource short name	n/a
Landing page	https://commoncrawl.org/
Description	Open repository of web crawl data. The Common Crawl corpus contains petabytes of data collected since 2008. It contains raw web page data, extracted metadata and text extractions. The corsican part contains 75 944 pages. Data available under specific licence.
(Funding) project	n/a
Language(s)	Corsican (amongst many languages)
Resource publication year	2021 (April)
Corpus subclass	raw corpora
Media type(s) of parts	texts

2.9. OPUS Wikimedia

Resource name	OPUS Wikimedia
Resource type	corpus
Ressource short name	n/a
Landing page	https://opus.nlpl.eu/wikimedia.php
Description	This corpus provides parallel data based on Wikimedia content. The Corsican data includes a set of tokenised sentences (around 36K tokens) and aligned data with several languages (Catalan, English, Esperanto, Spanish, French, Galician, Italian, Latin, Potuguese, Chinese and Zulu). License: CC–BY-SA 4.0
(Funding) project	n/a
Language(s)	Corsican (and Catalan, English, Esperanto, Spanish, French, Galician, Italian, Latin, Potuguese, Chinese, Zulu)
Resource publication year	2019 -2021
Corpus subclass	annotated corpora
Media type(s) of parts	texts

2.10. WikiANN

Resource name	WikiANN
Resource type	corpus
Ressource short name	n/a
Landing page	https://elisa-ie.github.io/wikiann/
Description	WikiANN is a multilingual named entity recognition dataset consisting of Wikipedia articles annotated with LOC (location), PER (person), and ORG (organisation) tags in the IOB2 format. This data is licensed under the Attribution License (ODC-By). For research use only. There is a balanced train, dev, and test splits of Rahimi et al. (2019), which supports 176 of the 282 languages from the original WikiANN corpus : https://huggingface.co/datasets/viewer/?dataset=wikiann&config=co
(Funding) project	n/a
Language(s)	Corsican (amongst 282 languages)
Resource publication year	2017
Corpus subclass	annotated corpora
Media type(s) of parts	texts

2.11. Multilingual C4

Resource name	Multilingual C4
Resource type	corpus
Ressource short name	mC4
Landing page	https://www.tensorflow.org/datasets/catalog/c4#c4multilingual
Description	A colossal, cleaned version of Common Crawl's web crawl corpus. Based on Common Crawl dataset (https://commoncrawl.org). Data is not available as is but must be reconctructed following the provided instructions. Licensed under CC BY 4.0.
(Funding) project	Google (TensorFlow)
Language(s)	Corsican (amongst 101 languages)
Resource publication year	2021
Corpus subclass	raw corpora
Media type(s) of parts	texts

2.12. Miriadi Lectŭrĭo+

Resource name	Miriadi Lectŭrĭo+
Resource type	corpus
Ressource short name	Lecturio+
Landing page	https://www.miriadi.net/lecturio
Description	This project offers a corpus of stories for very young children, translated into several languages and sometimes into several versions. Contents under CC BY-NC-ND 3.0 licence.
(Funding) project	Erasmus+ (European Union)
Language(s)	Corsican (amongst other languages : French, Spanish, German, Catalan, Italian, Portuguese, English, Occitan, Romanian, Armenian)
Resource publication year	2017-2019
Corpus subclass	raw corpora
Media type(s) of parts	texts

2.13. Pruverbii, detti è sprissioni di Corsica è d’altrò

Resource name	Pruverbii, detti è sprissioni di Corsica è d’altrò
Resource type	corpus
Ressource short name	n/a
Landing page	http://ernestpapi.free.fr/
Description	Proverbs, sayings and expressions from Corsica and elsewhere.
(Funding) project	n/a
Language(s)	Corsican (Italian and other languages)
Resource publication year	?
Corpus subclass	raw corpora
Media type(s) of parts	texts

3. Oral Corpora

3.1. Cocoon, the "Enquête en Corse" collection

Resource name	**Cocoon, the "Enquête en Corse" collection**
Resource type	corpus
Ressource short name	n/a
Landing page	https://cocoon.huma-num.fr/exist/crdo/meta/cocoon-040a0841-6ea2-35bf-bdbd-aa660385c8c0 or https://doi.org/10.34847/cocoon.040a0841-6ea2-35bf-bdbd-aa660385c8c0
Description	A collection of native corsican speakers audio recordings, sometimes with transcripts (corsican+french translation), on various topics :Olive growing in Corsica ; The economy of a Corsican village in the 1950s ; Bread making in the past in Corsica ; Wine growing in Corsica ; Honey and bees in Corsica ; Goats and sheep ; Porto-Vecchio in the past seen by a villager. Data under CC BY-NC-ND 2.5 licence. • L'oléiculture en Corse 1 (audio) • L'oléiculture en Corse 2 (audio) • L'oléiculture en Corse 3 (audio+transcript) • L'oléiculture en Corse 4 (audio+transcript) • L'oléiculture en Corse 5 (audio+transcript) • L'oléiculture en Corse 6 (audio+transcript) • L'oléiculture en Corse 7 (audio) • L'oléiculture en Corse 8 (audio) • L'oléiculture en Corse 9 (audio) • L'oléiculture en Corse 10 (audio+transcript) • L'économie d'un village corse dans les années 50 1 (audio+transcript) • L'économie d'un village corse dans les années 50 2 (audio+transcript) • L'économie d'un village corse dans les années 50 3 (audio+transcript)

	• L'économie d'un village corse dans les années 50 4 (audio+transcript) • L'économie d'un village corse dans les années 50 5 (audio+transcript) • L'économie d'un village corse dans les années 50 6 (audio) • L'économie d'un village corse dans les années 50 7 (audio) • L'économie d'un village corse dans les années 50 8 (audio) • La fabrication du pain autrefois en Corse (audio) • La viticulture en Corse 1 (audio) • La viticulture en Corse 2 (audio) • La viticulture en Corse 3 (audio) • La viticulture en Corse 4 (audio) • La viticulture en Corse 5 (audio) • La viticulture en Corse 6 (audio) • La viticulture en Corse 7 (audio) • La viticulture en Corse 8 (audio) • La viticulture en Corse 9 (audio+transcript) • Le miel et les abeilles en Corse 1 (audio) • Le miel et les abeilles en Corse 2 (audio) • Pecure e capre 1 (audio) • Pecure e capre 2 (audio+transcript) • Pecure e capre 3 (audio) • Pecure e capre 4 (audio) • Porto-Vecchio autrefois vu par un villageois 1 (audio+transcript) • Porto-Vecchio autrefois vu par un villageois 2 (audio+transcript) • Porto-Vecchio autrefois vu par un villageois 3 (audio+transcript) • Porto-Vecchio autrefois vu par un villageois 4 (audio+transcript) • Porto-Vecchio autrefois vu par un villageois 5 (audio+transcript) • Porto-Vecchio autrefois vu par un villageois 6 (audio+transcript)
(Funding) project	Banque de Données Langue Corse (BDLC), Université de Corse, CNRS
Language(s)	Corsican, French (transcript)

Resource publication year	2015
Corpus subclass	raw corpora, annotated corpora
Media type(s) of parts	texts, audio

3.2. Speaking Atlas of the Regional Languages of France

Resource name	Speaking Atlas of the Regional Languages of France
Resource type	corpus
Ressource short name	n/a
Landing page	http://catalog.elra.info/en-us/repository/browse/ELRA-S0402/
Description	The Speaking atlas of the regional languages of France offers the same Aesop's fable read in French and in a number of varieties of languages of France. This work, which has a scientific and heritage dimension, consists in highlighting the linguistic diversity of Metropolitan France and Overseas Territories, through recordings collected in the field and presented via an interactive map, with their orthographic transcription. As far as Occitan is concerned, about sixty varieties were collected in Gascony, Languedoc, Provence, northern Occitania and the Linguistic Crescent. Varieties of Basque, Breton, Franconian, West Flemish, Alsatian, Corsican, Catalan, Francoprovençal and Oïl language(s) are also provided, as well as about fifty languages in the French Overseas and non-territorial languages such as Rromani and the French sign langage. Data available under CC-BY-NC-SA licence.
(Funding) project	n/a
Language(s)	Corsican (+other regional langage of France)
Resource publication year	2018
Corpus subclass	raw corpora
Media type(s) of parts	texts, audio

4. Tools ans applications

4.1. Compact Language Detector 2

Resource name	**Compact Language Detector 2**
Resource type	tool / service
Ressource short name	CLD2
Landing page	https://github.com/CLD2Owners/cld2; https://pypi.org/project/pycld2/ (python binding)
Description	CLD2, Compact Language Detector 2, is the language detection component of Google's Chromium browser. This system is based on a naive Bayesian approach using ngrams. It is a module available on the Github platform that is documented as being able to recognise 83 languages. The latest version however supports a larger set of 161 languages, including Corsican. Tool available under the Apache v.2 licence. Also available as a python binding (https://pypi.org/project/pycld2/).
(Funding) project	Google
Language(s)	161 languages, including Corsican
Resource publication year	2013-2015
Media type(s) of input and output	text text
Function(s) / Task(s)	Language detection/identification
Language dependent	yes
TRL [1-9]	9 (actual system proven in operational environment)

4.2. Compact Language Detector 3

Resource name	Compact Language Detector 3
Resource type	tool / service
Ressource short name	CLD3
Landing page	https://github.com/google/cld3 https://github.com/bsolomon1124/pycld3
Description	CLD3 is a neural network tool for language identification. This package contains the inference code and a trained model for 213 languages, including Corsican. Tool available under the Apache v.2 licence. Also available as a python binding (https://github.com/bsolomon1124/pycld3).
(Funding) project	Google
Language(s)	213 languages, including Corsican
Resource publication year	2020
Media type(s) of input and output	text text
Function(s) / Task(s)	Language detection/identification
Language dependent	yes
TRL [1-9]	9 (actual system proven in operational environment)

4.3. FastText Language Identification

Resource name	FastText Language Identification
Resource type	tool / service
Ressource short name	n/a
Landing page	https://fasttext.cc/docs/en/language-identification.html
Description	FastText is an open-source, free, lightweight library that allows users to learn text representations and text classifiers. It offers two models for language identification, which can recognize 176 languages, including Corsican. These models were trained on data from Wikipedia, Tatoeba and SETimes, used under CC-BY-SA. FastText is available under MIT licence, models under CC BY-SA 3.0.
(Funding) project	Facebook
Language(s)	176 languages, including Corsican
Resource publication year	2016
Media type(s) of input and output	text text
Function(s) / Task(s)	Language detection/identification
Language dependent	yes
TRL [1-9]	8 (system complete and qualified)

4.4. Language-Aware String Extractor

Resource name	**Language-Aware String Extractor**
Resource type	tool / service
Ressource short name	La-strings Whatlang
Landing page	https://sourceforge.net/projects/la-strings/
Description	A language identification library based on the K-nearest neighbour algorithm and the cosine similarity measure. Version 1.25 with Language Data Release 4 supports 1547 languages, including Corsican. This tool is also known in the scientific literature as Whatlang Available under the GPLv.3 licence.
(Funding) project	n/a
Language(s)	~ 1500 languages, including Corsican
Resource publication year	2015 (v 1.25) 2020 (Language Data)
Media type(s) of input and output	text text
Function(s) / Task(s)	Language detection/identification
Language dependent	yes
TRL [1-9]	4 (technology validated in lab)

4.5. Yet Another Language Identifier

Resource name	Yet Another Language Identifier
Resource type	tool / service
Ressource short name	YALI
Landing page	https://ufal.mff.cuni.cz/tools/yali
Description	YALI is tool for language identification with pretrained models for 122 languages. Available as a Perl CPAN module Lingua::YALI. Tool available under the BSD licence.
(Funding) project	n/a
Language(s)	Corsican (amongst 122 languages)
Resource publication year	2012, last update : 2019
Media type(s) of input and output	text text
Function(s) / Task(s)	Language detection/identification
Language dependent	yes
TRL [1-9]	4 (technology validated in lab)

4.6. Okchakko Translator

Resource name	Okchakko Translator
Resource type	tool / service
Ressource short name	n/a
Landing page	https://okchakko.com/ Demo : https://www.okchakko.com/translate/ API : https://rapidapi.com/okchakkotranslator/api/okchakko-translator/details
Description	Translate from French to Corsican language (3 dialectal variations : cismuntincu, sartinese, taravese). Available as an online demontration, Windows application, Android app and API (freemium model).
(Funding) project	n/a
Language(s)	Corsican, French
Resource publication year	2020
Media type(s) of input and output	text text
Function(s) / Task(s)	Computer-aided translation
Language dependent	yes
TRL [1-9]	3 (experimental proof of concept)