THE MULTIDIALECTAL CORPUS OF THE CRESCENT DIALECTS

COLLECTION, EXPLOITATION AND ANALYSIS

This work is part of the following projects, both funded by the French National Research Agency: ANR-17-CE27-0001-01 (Project "The Linguistic Crescent: A Multidisciplinary Approach to a Contact Area between Oc and Oïl varieties") & ANR-10-LABX-0083 (program "Investissements d'Avenir", Labex EFL, Strand 3, Workpackage VC2 - "Central Gallo-Romance: linguistics and ecology of a transitional zone"), and the project "Oc/Oïl: texts, identity and language contact" funded by the City of Paris.

Maximilien GUÉRIN
Postdoctoral Researcher
Université Paris Cité & CNRS HTL (UMR 7597)
mguerin.ling@gmail.com

Corpus Building: Linguistic Crescent

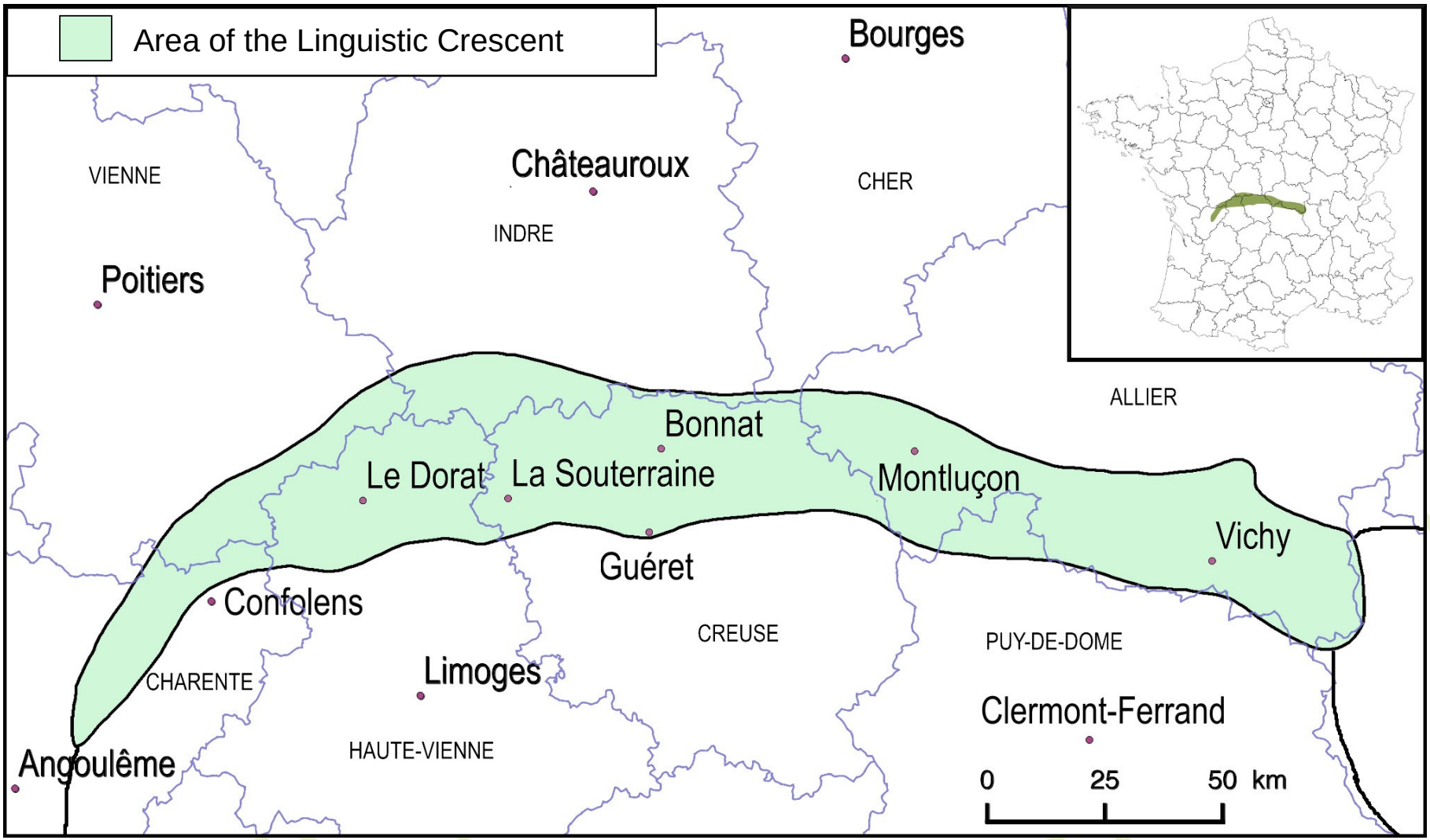
- Situated on the Northern Fringe of the Massif Central (France)
  - Form of a half moon (crescent)

- Local gallo-romance dialects simultaneously display typical features:
  - Oïl varieties (French, Poitevin-Saintongeais, Berrichon…)
  - Oc varieties (Limousin (west) and Auvergnat (east))

- Crescent dialects often called
  - ‘Marchois’ (~ ‘Marchese’) - Limousin side
  - ‘Bourbonnais’ - Auvergnat side

- Crescent area = dialect continuum
  - Variation very great (especially north-south axis)
  - More than 20km → intercomprehension may be difficult

Corpus Building: Linguistic Crescent



Corpus Building: Population

- Crescent = endangered languages
  - Nearly all speakers = bilingual since 19th Century
  - Remained vernacular language until middle 20th century
  - After WW2, “process of disappearing”
    = breaking line generational transmission

- Languages of oral tradition
  - No written literature (except some recent texts)
  - No literary or standard language

- Each one speak her/his own dialect
  - Few existing texts written in specific dialects

- Typology of speakers:
  > 70 yr. = native and fluent speakers
  40-70 yr. = terminal speakers (some knowledge/skills)
  < 40 yr. = French monolinguals

Corpus Building: Initial aims

- Save what can be saved
  - In 2040: no more native speakers

- Why make this corpus?
  - Make the corpus accessible to the local population
  - Make a corpus useful for research:
    ~ Romance linguistics
    ~ Typology
    ~ Sociolinguistics
    ~ Natural language processing
    ~ etc.

Corpus Building: Data collection

- Fieldwork (> 70 towns/villages)

  https://www.google.com/maps/...

Corpus Building: Questionnaires

- Linguistic questionnaires

  http://tulquest.huma-num.fr/en & https://parlersducroissant.huma-num.fr/participer.html

  - Basic lexicon - Nouns
  - Basic lexicon - Pronouns
  - Basic lexicon - Other
  - Additional lexicon
  - Lexicon - Calendar time
  - Conjugation

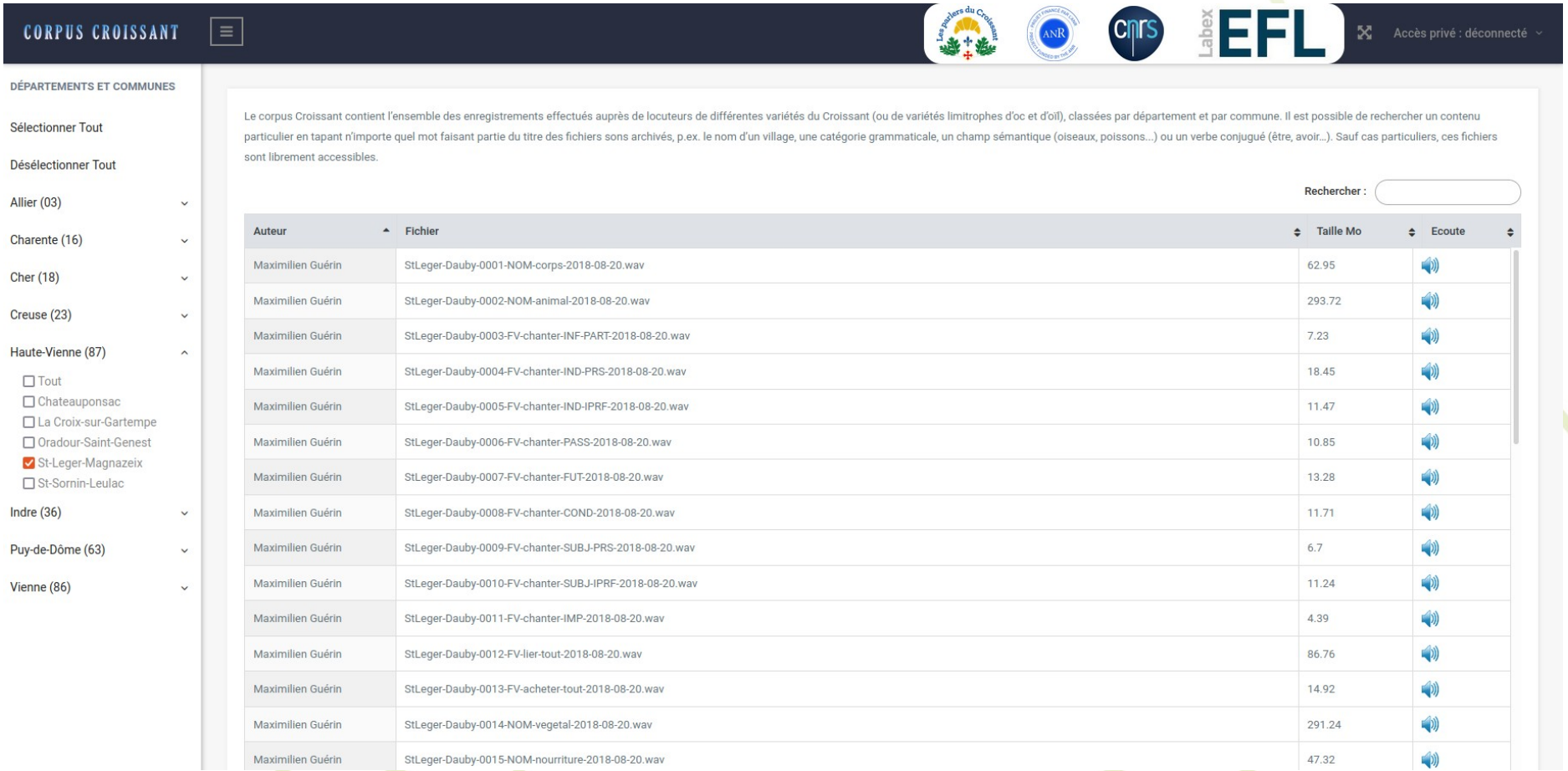
  ↳ Fieldwork

1. Parties du Corps		
	1.1. Tête	
1	Barbe	
2	Bouche	
3	Cheveu	
4	Cil	
5	Cou	
6	Dent	
7	Front	
8	Gosier	
9	Joue	
10	Langue	
11	Lèvre	
12	Menton	
13	Molaire	
14	Moustache	
15	Nez	
16	Nuque	
17	Œil	
18	Oreille	
19	Paupière	
20	Pupille de l'œil	
21	Sourcil	

Corpus Building: Oral part

- Corpus Crescent: lexicon and morphology

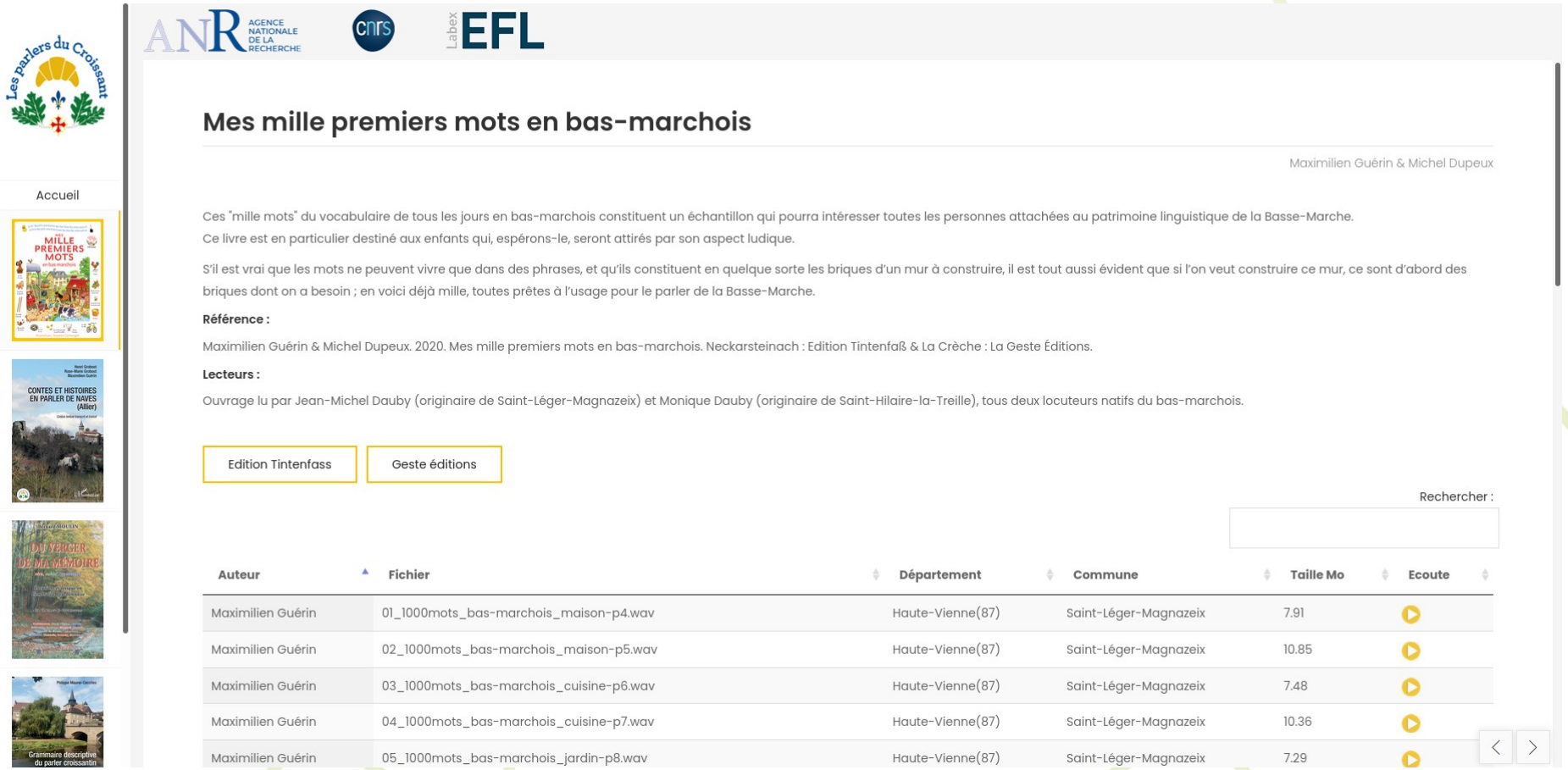
  https://parlersducroissant.huma-num.fr/corpus/



Corpus Building: Oral part

- Audiobooks Crescent: texts

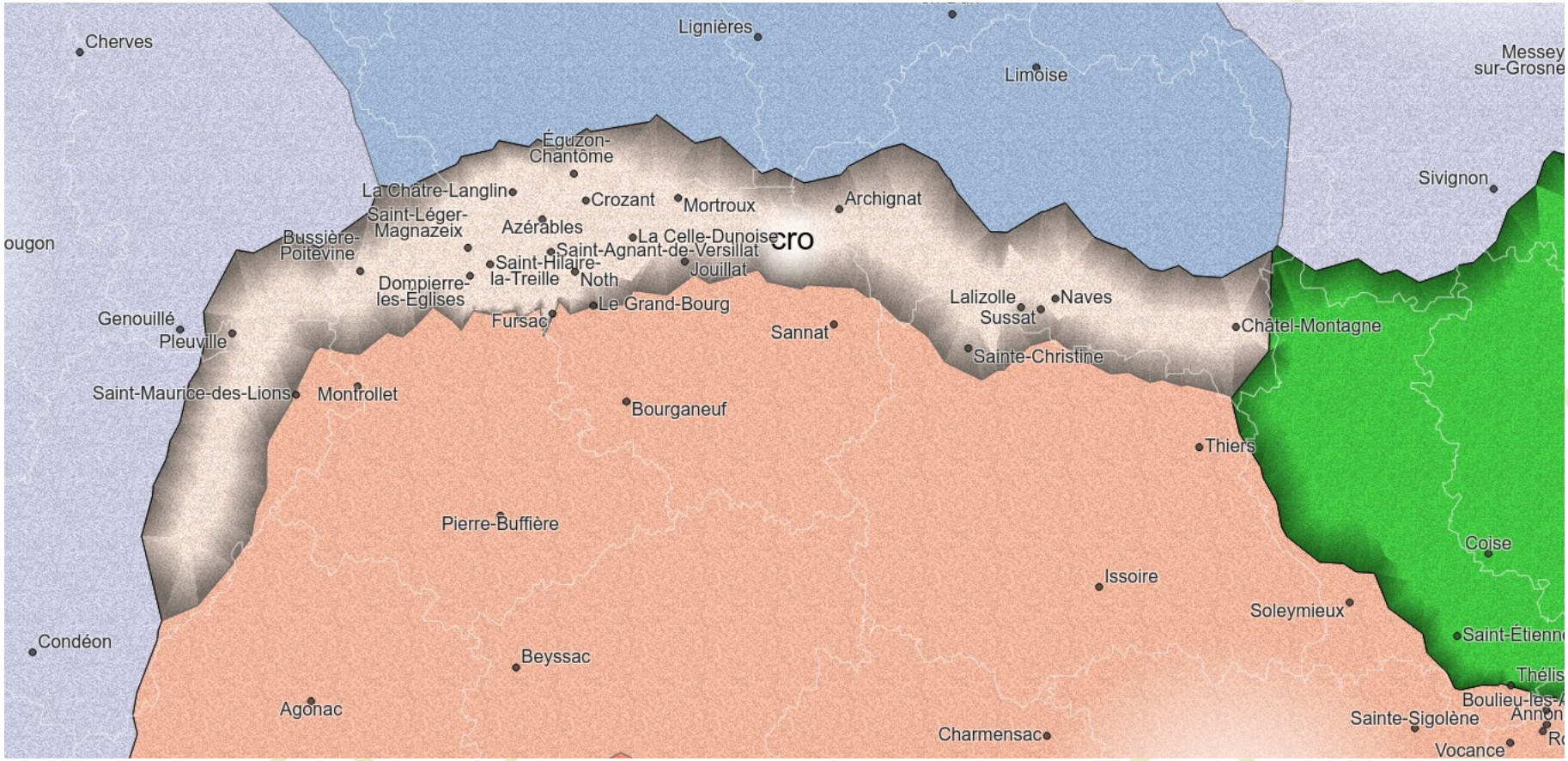
  https://parlersducroissant.huma-num.fr/livres-audios/



Corpus Building: Oral part

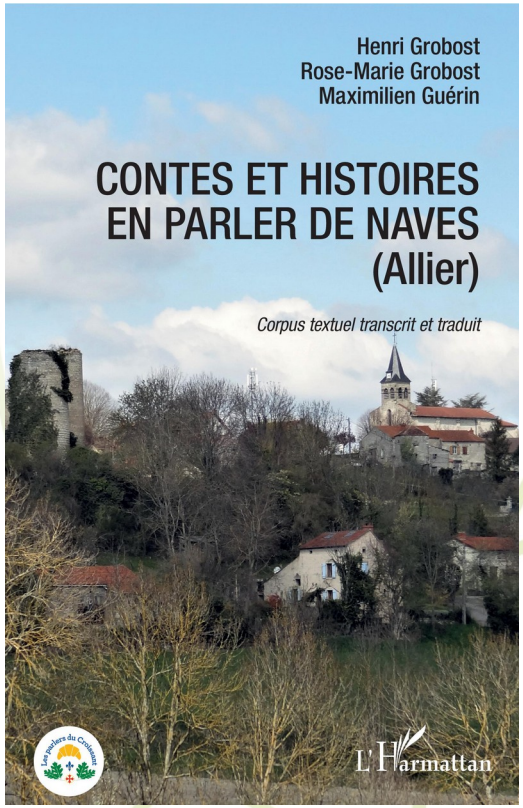
- Speaking atlas of the regional languages of France

  https://atlas.limsi.fr/?tab=cro



Corpus Building: Written part

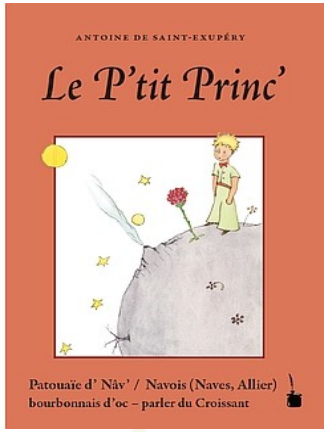
- Text corpus
  - Orthographic transcription
  - IPA transcription
  - Translation

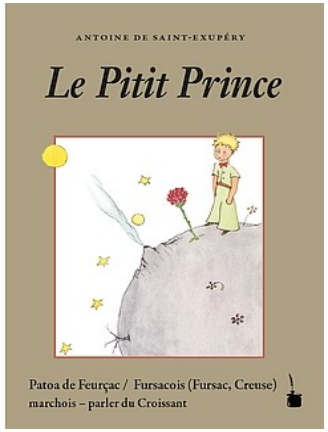


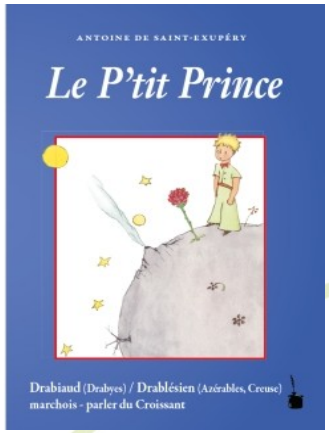
- Text corpus
  - "Historical" corpus
  - Orthographic transcription
  - Translation

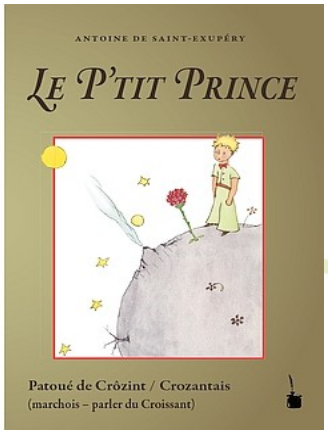
Corpus Building: Written part

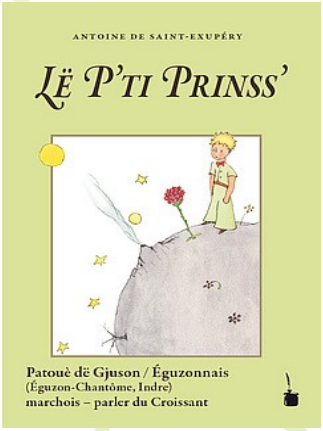
- Translations of “The Little Prince”
  - Typological parallel corpus (cf. Stolz 2009)

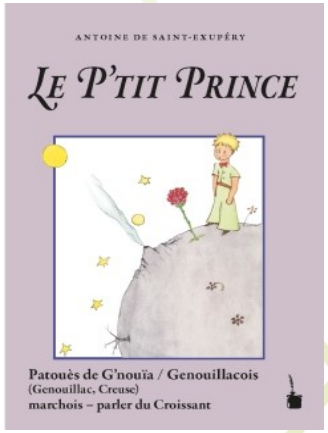


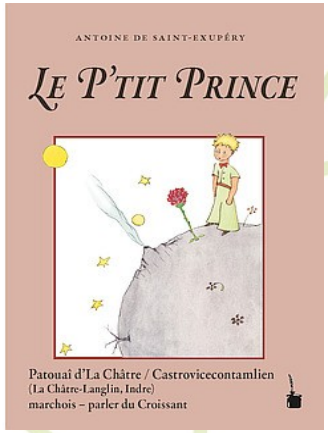


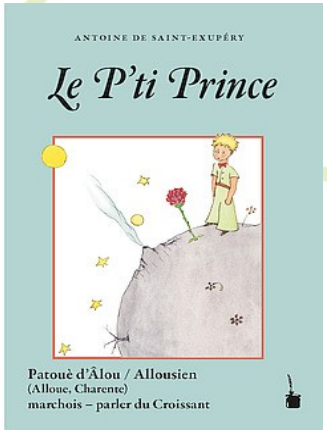


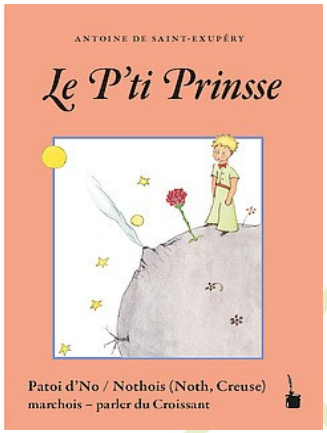


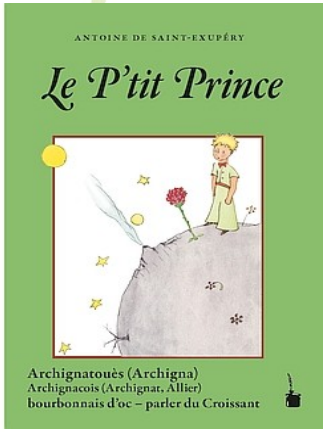












Corpus Building: Metadata

- Informant
  - Name
  - Date of birth / Age
  - Gender
  - Place of birth
  - Places of residence
  - Education level
  - Occupation
  - Family situation
  - Mother tongue
  - Home language

↳ Useful for sociolinguistic studies

MÉTADONNÉES POUR
LES ENREGISTREMENTS

LES PARLERS DU CROISSANT

PROJETS PORTÉS PAR LES LABORATOIRES LLACAN (CNRS - UMR 8135) ET LIMSI (CNRS - UPR 3251),
FINANCÉS PAR L'ANR (ANR-17-CE27-0001-01) ET LE LABEX EFL (AXE 3, OPÉRATION LC4)
SOUS LA DIRECTION DE NICOLAS QUINT, DIRECTEUR DE RECHERCHE AU CNRS

Date : ……/……/…… Nom et prénom de l'enquêteur : .....

Département : .....

Commune (nom et code postal) : .....

Village (hameau, lieu-dit) : .....

Informateur :

Nom : ..... Prénom : ..... Sexe : H / F

Date de naissance : ……/……/…… Âge (lors de l'enregistrement) : .....

Lieu de naissance : .....

Lieu de résidence actuel : .....

Lieu de résidence successifs	Période (années)

Niveau d'étude : .....

Profession actuelle (si retraité, profession avant retraite) : .....

Professions successives	Période (années)

Situation familiale (marié/e, divorcé/e, etc.) : ..... Enfants (nombre) : .....

Langues parlées : .....

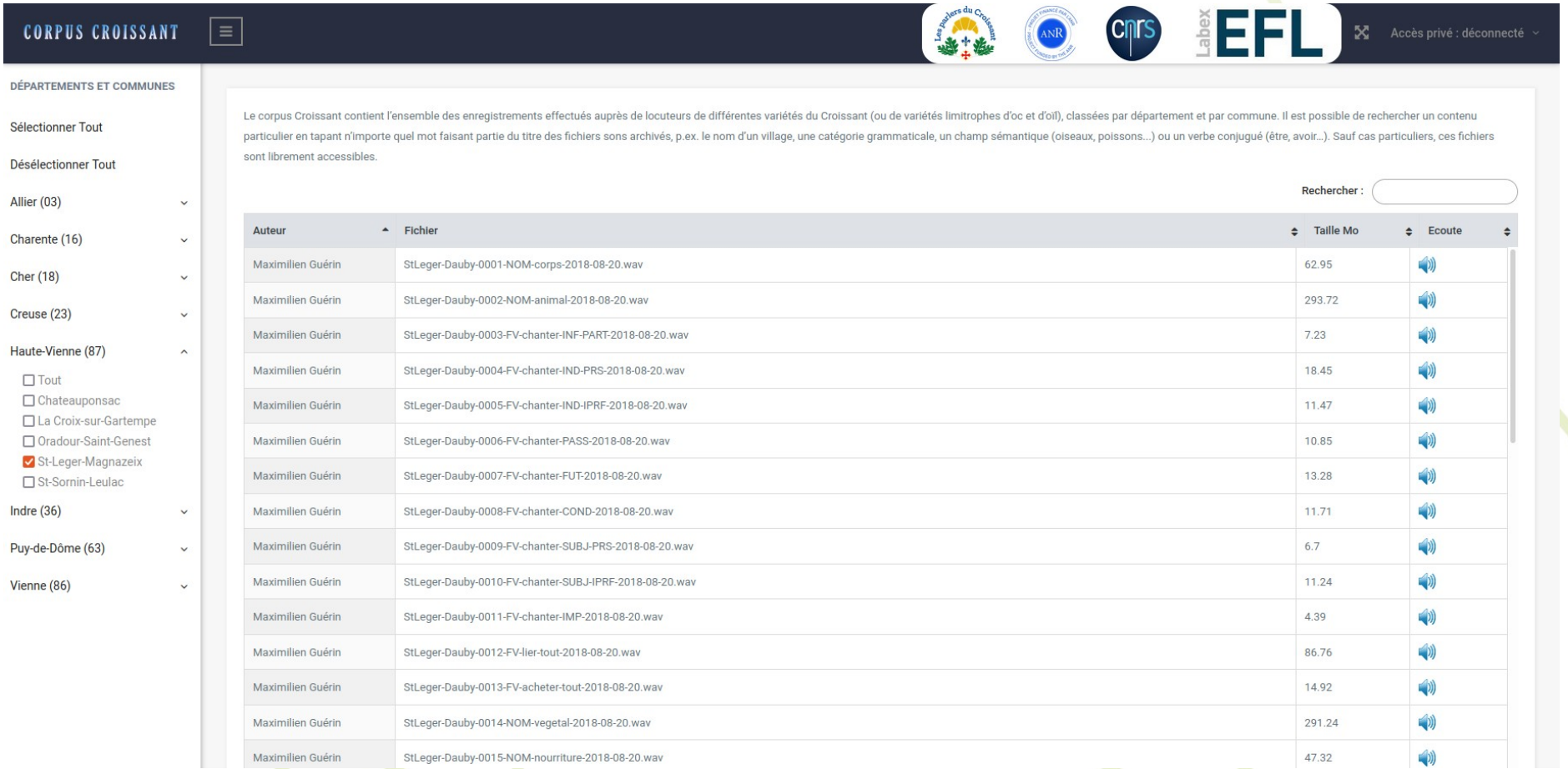
Première langue (pratiquée pendant l'enfance) : ..... Langue de la maison (pendant l'enfance) : .....

Annotation of the corpus

- Pragmatic choise: priority to collection
  - There will soon be no more speakers
- Organization of the morphological part
- Texts to annotate

Exports and data management

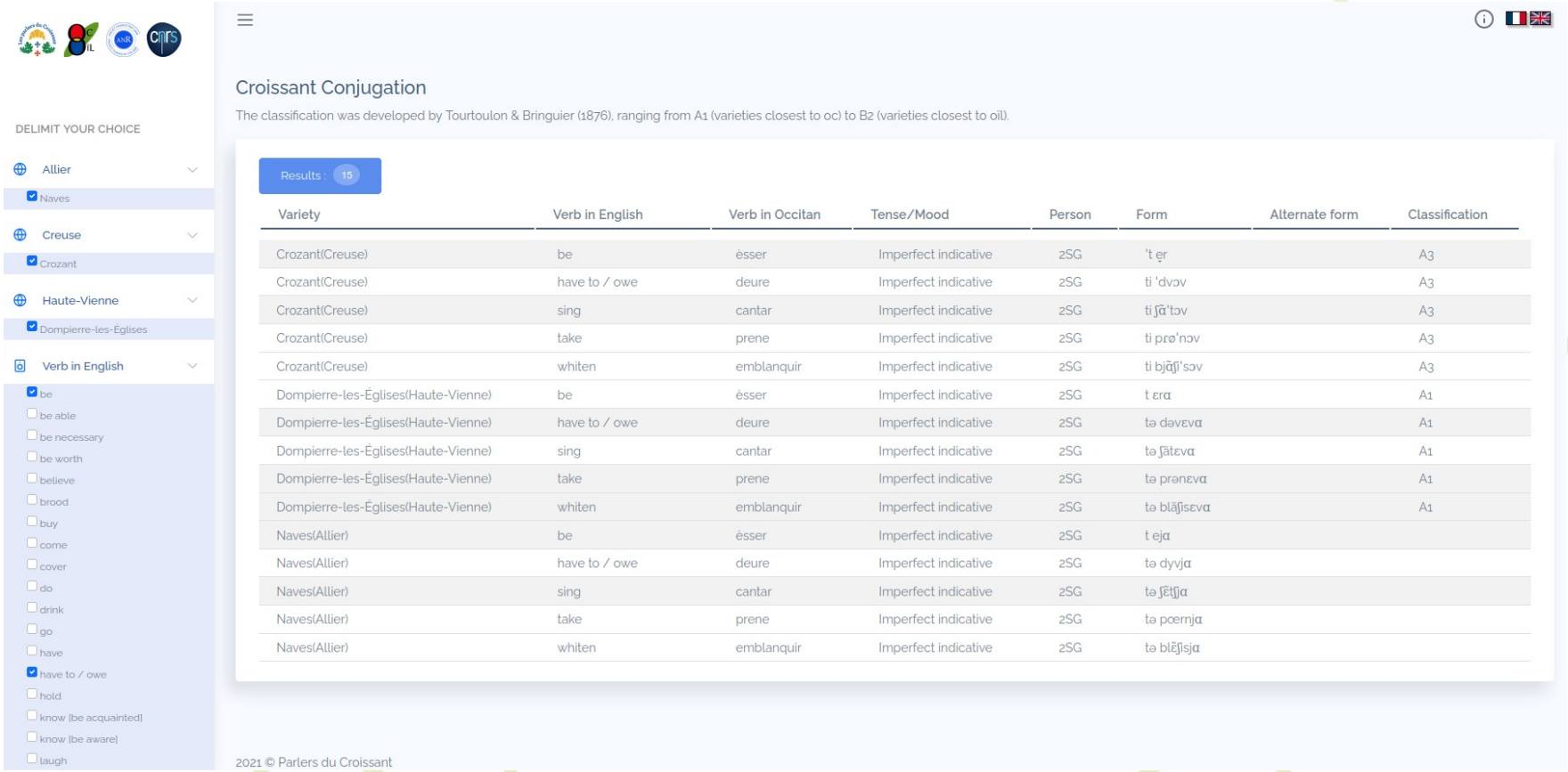
- Mainly raw/primary data
- Oral data: .wav files (44kHz - 24bits)



Exports and data management

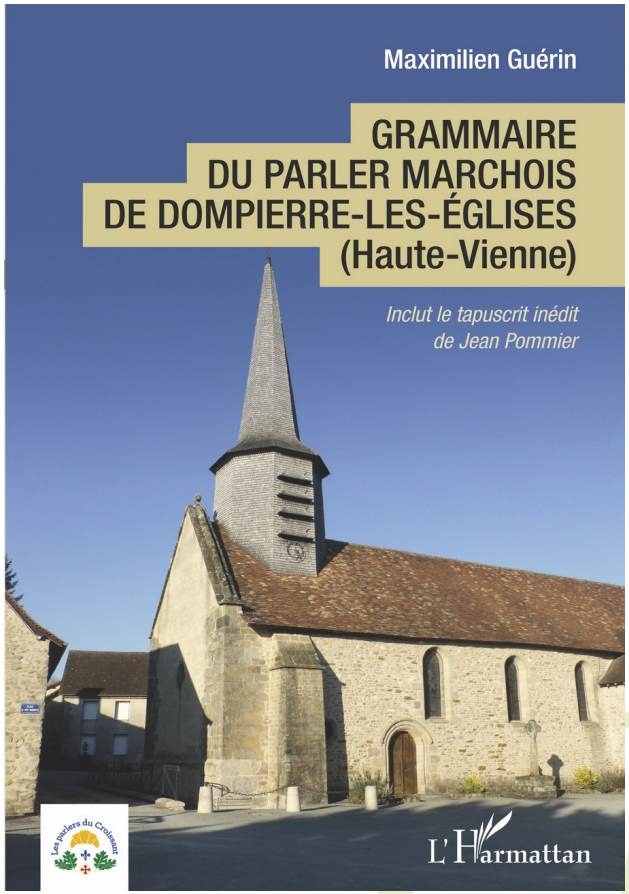
- Target levels of analysis: phonology, morphology
- To develop: semantics, syntax

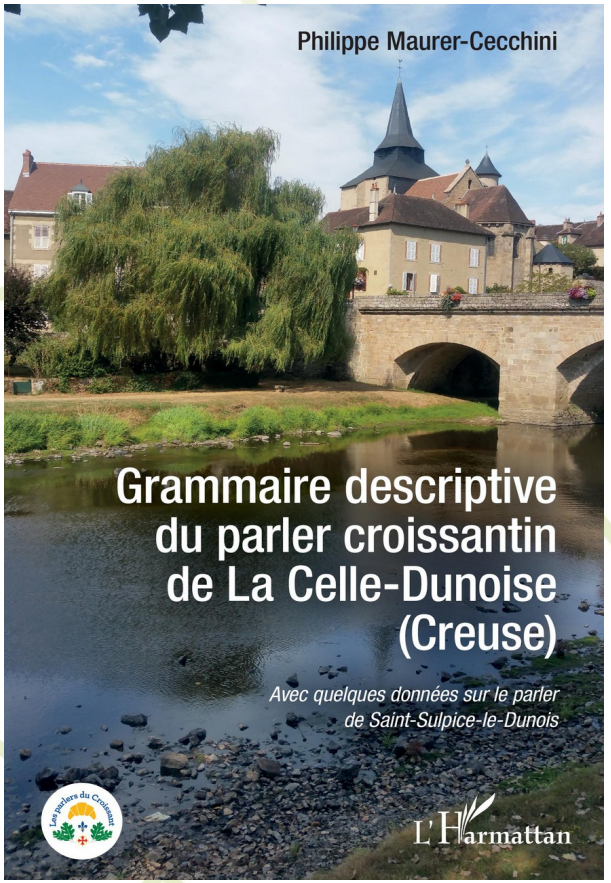
https://parlersducroissant.huma-num.fr/conjugaison/



Analysis: Examples of the results

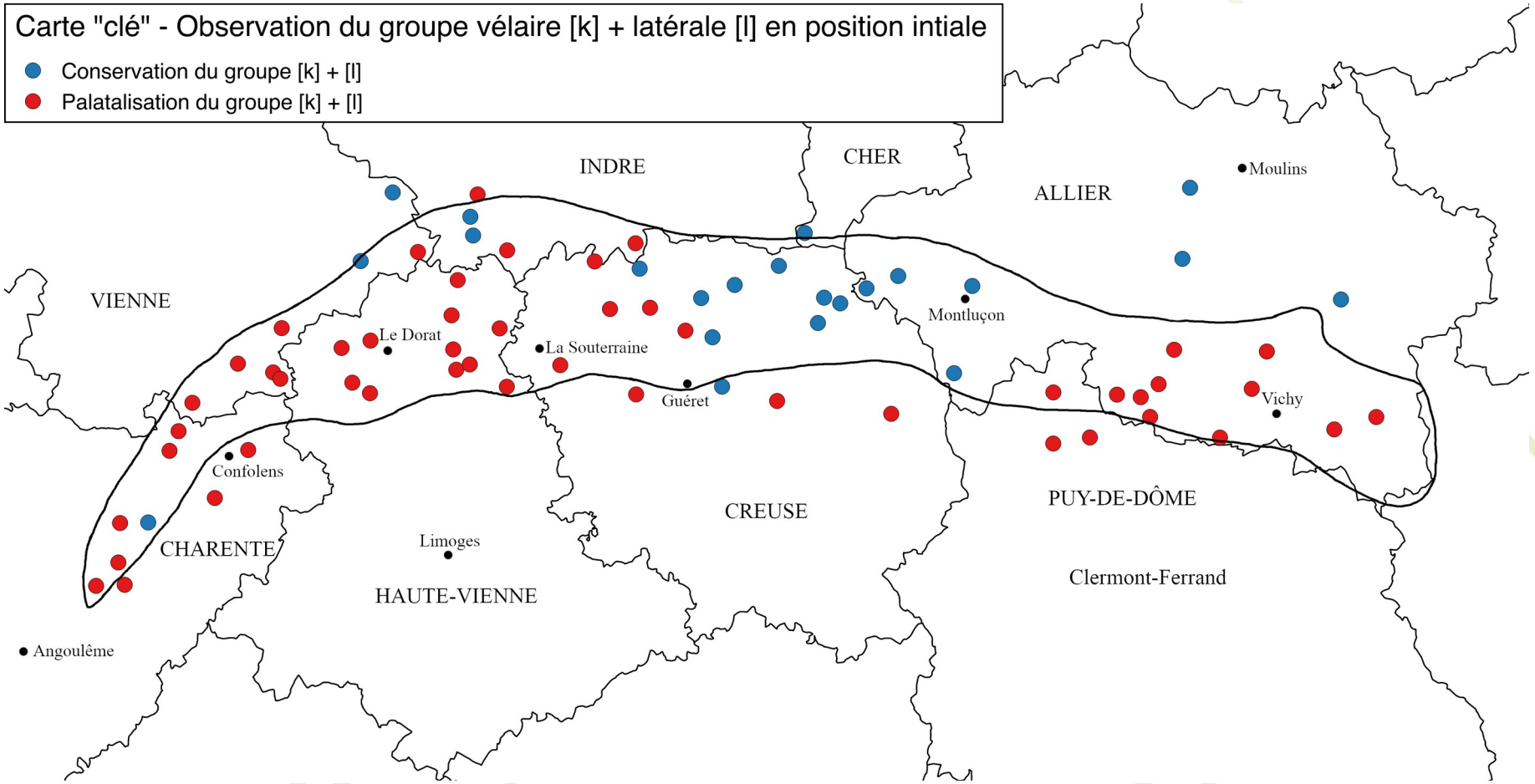
- Grammatical descriptions





Analysis: Examples of the results

- Dialectical maps – Phonological analysis

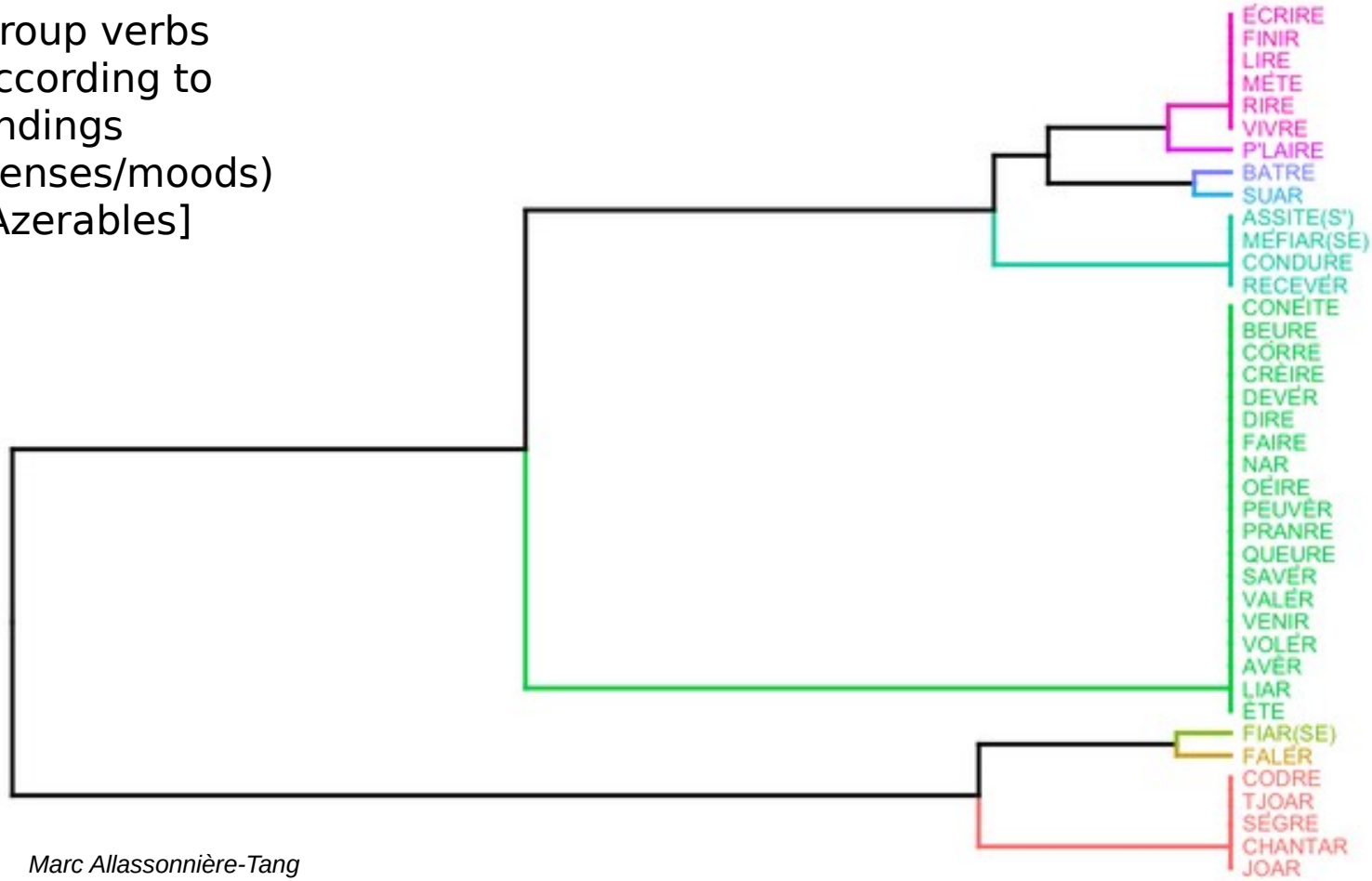


Amélie Deparis

Analysis: Examples of the results

- Hierarchical clustering – Morphological analysis

Group verbs according to endings (tenses/moods) [Azerables]

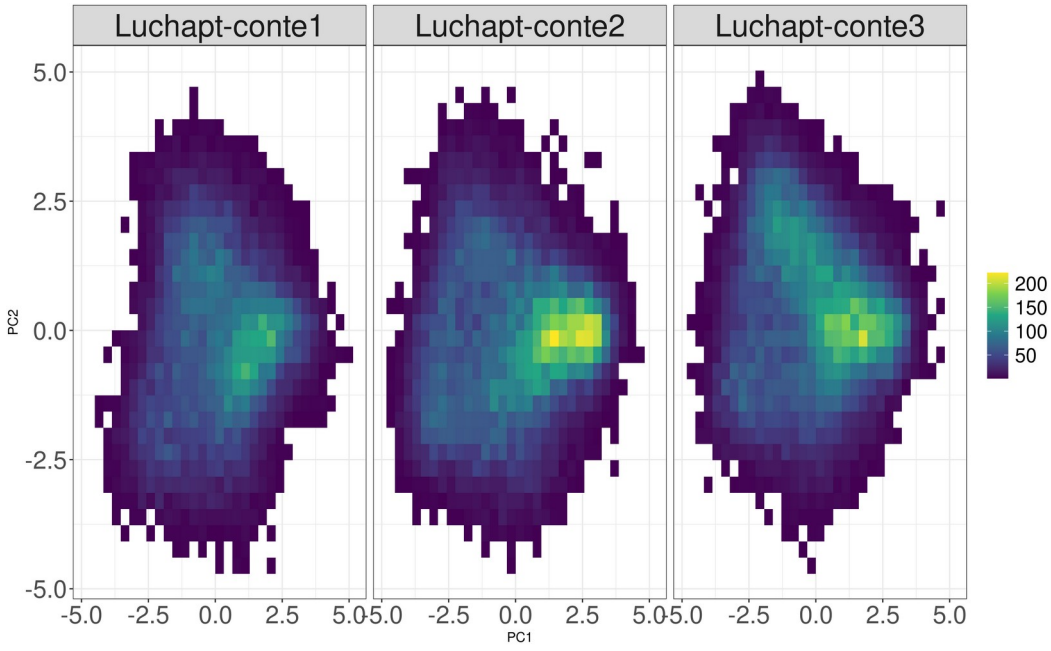


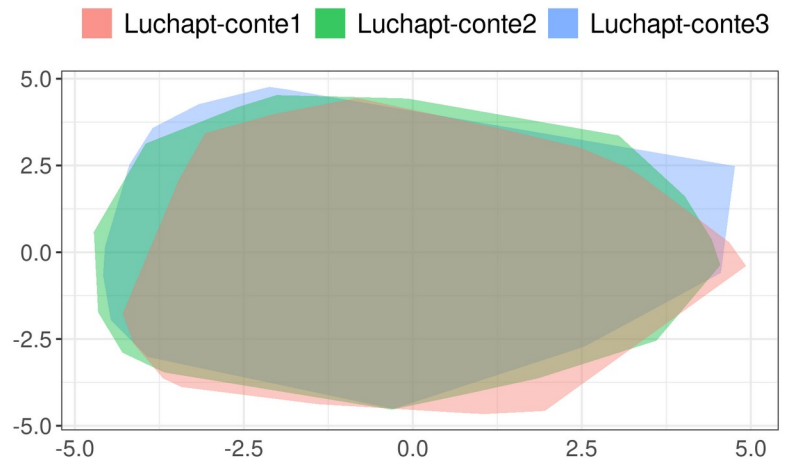
Marc Allassonnière-Tang

Analysis: Examples of the results

- Mel-frequency cepstral coefficients → Acoustic space

  Phonetic comparison





Marc Allassonnière-Tang

Discussion

- Initial aims vs. Results
  - Achieved aims
  - Good reactions from the local population
  - Used for research works

- Implications for language sciences
  - Several data for under-described dialects
  - Large typological parallel corpus
  - New elements for romance linguistics

Limitations

- Orthographic choices
  - Not the same conventions for all texts
- Phonetic transcriptions
  - Some phonetic issues to solve
- Some parts of the area
  - Difficulties to find speakers in some places
- Enormous amount to do

Future perspectives

- Continue to develop our corpora
  - Find new places
  - More “The Little Prince” translations
  - More audiobooks
  - More online tools
  - etc.
- Do annotation work
- More durable archive
  - Cocoon platform (Digital Oral Corpus COllections)
- Discussions about speakers’ and authors’ rights

References

- Esher Louise, Guérin Maximilien, Quint Nicolas, Russo Michela. (eds). 2021. *Le Croissant linguistique : entre oc, oïl et franco-provençal - Des mots à la grammaire, des parlers aux aires*. Paris: L'Harmattan.
- Grobost Henri, Grobost Rose-Marie, Guérin Maximilien. 2020. *Contes et histoires en parler de Naves (Allier) : Corpus textuel transcrit et traduit*. Paris: L'Harmattan.
- Guérin Maximilien, Meddour Tahar & Quint Nicolas (eds.). 2018. *Corpus Croissant*. Villejuif: LLACAN, https://parlersducroissant.huma-num.fr/corpus/ (Online).
- Knyazeva Elena, Adda Gilles, Boula de Mareüil Philippe, Guérin Maximilien & Quint Nicolas. 2020. Automatic Extraction of Verb Paradigms in Regional Languages: the case of the Linguistic Crescent varieties. In Dorothee Beermann et al. (éds.), *Proceedings of the LREC 2020 1st Joint SLTU and CCURL Workshop (SLTU-CCURL 2020)*, 245-249. Paris: European Language Resources Association (ELRA), http://www.lrec-conf.org/proceedings/lrec2020.
- Rémy Marcel. 2021. *Patoiseries de "La Soutrane" : Edition préfacée et traduite par Maximilien Guérin*. Paris: L'Harmattan.
- Stolz Thomas. 2009. *Harry Potter* meets *Le petit prince* – On the usefulness of parallel corpora in crosslinguistic investigations. *Sprachtypologie und Universalienforschung* 60(2), 100-117.