

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.



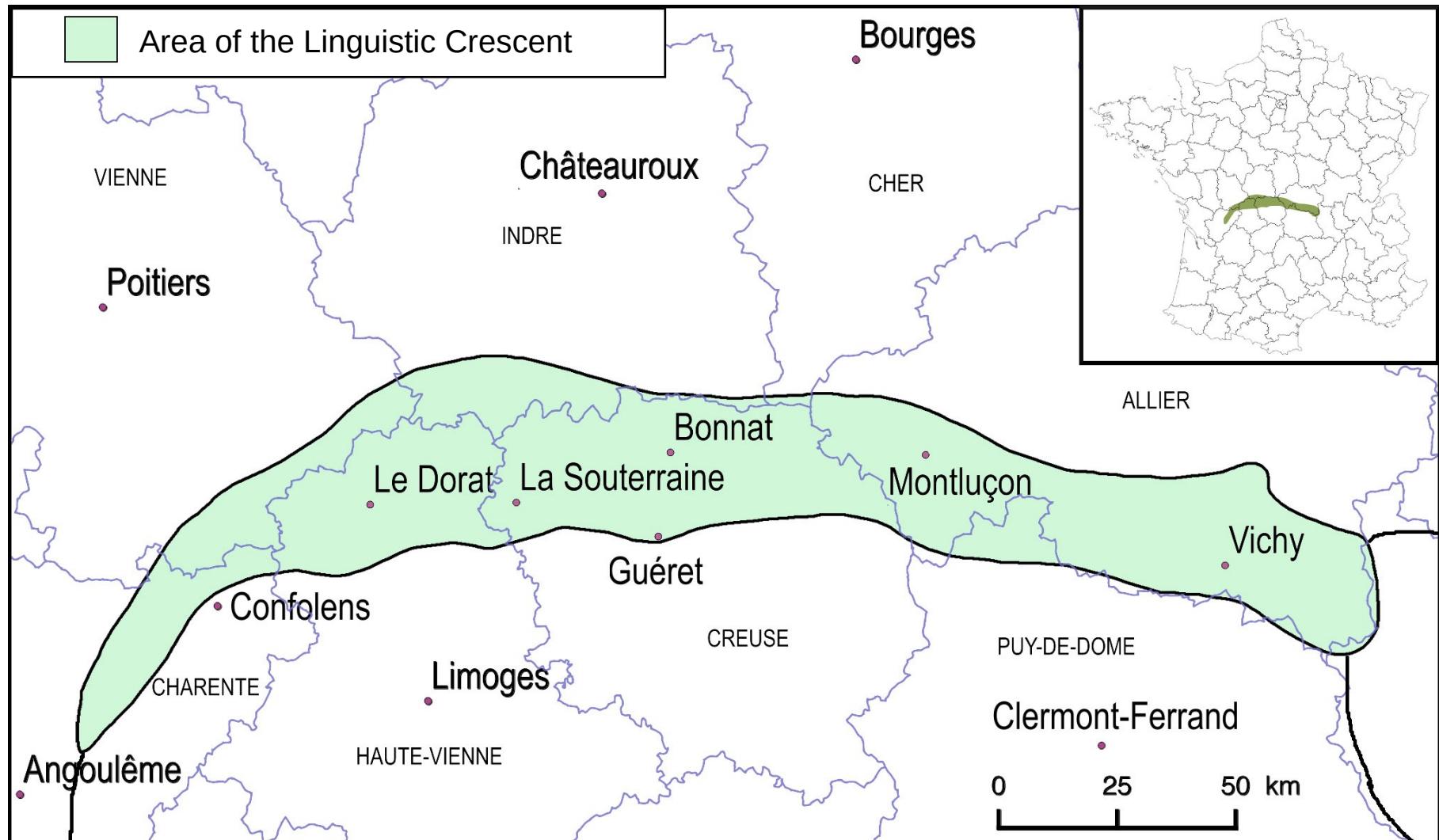
Maximilien GUÉRIN 
Postdoctoral Researcher
Université Paris Cité & CNRS HTL (UMR 7597)
mguerin.ling@gmail.com



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/...](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
22	

Corpus Building: Oral part

- Corpus Crescent: lexicon and morphology

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

CORPUS CROISSANT ☰

DÉPARTEMENTS ET COMMUNES

Selectionner Tout

Désélectionner Tout

Allier (03)

Charente (16)

Cher (18)

Creuse (23)

Haute-Vienne (87)

Tout

Chateauponsac

La Croix-sur-Gartempe

Oradour-Saint-Genest

St-Leger-Magnazeix

St-Sornin-Leulac

Indre (36)

Puy-de-Dôme (63)

Vienne (86)

Le corpus Croissant contient l'ensemble des enregistrements effectués auprès de locuteurs de différentes variétés du Croissant (ou de variétés limitrophes d'oc et d'oïl), classées par département et par commune. Il est possible de rechercher un contenu particulier en tapant n'importe quel mot faisant partie du titre des fichiers sons archivés, p.ex. le nom d'un village, une catégorie grammaticale, un champ sémantique (oiseaux, poissons...) ou un verbe conjugué (être, avoir...). Sauf cas particuliers, ces fichiers sont librement accessibles.

Rechercher :

Auteur	Fichier	Taille Mo	Ecoute
Maximilien Guérin	StLeger-Dauby-0001-NOM-corps-2018-08-20.wav	62.95	🔊
Maximilien Guérin	StLeger-Dauby-0002-NOM-animal-2018-08-20.wav	293.72	🔊
Maximilien Guérin	StLeger-Dauby-0003-FV-chanter-INF-PART-2018-08-20.wav	7.23	🔊
Maximilien Guérin	StLeger-Dauby-0004-FV-chanter-IND-PRS-2018-08-20.wav	18.45	🔊
Maximilien Guérin	StLeger-Dauby-0005-FV-chanter-IND-IPRF-2018-08-20.wav	11.47	🔊
Maximilien Guérin	StLeger-Dauby-0006-FV-chanter-PASS-2018-08-20.wav	10.85	🔊
Maximilien Guérin	StLeger-Dauby-0007-FV-chanter-FUT-2018-08-20.wav	13.28	🔊
Maximilien Guérin	StLeger-Dauby-0008-FV-chanter-COND-2018-08-20.wav	11.71	🔊
Maximilien Guérin	StLeger-Dauby-0009-FV-chanter-SUBJ-PRS-2018-08-20.wav	6.7	🔊
Maximilien Guérin	StLeger-Dauby-0010-FV-chanter-SUBJ-IPRF-2018-08-20.wav	11.24	🔊
Maximilien Guérin	StLeger-Dauby-0011-FV-chanter-IMP-2018-08-20.wav	4.39	🔊
Maximilien Guérin	StLeger-Dauby-0012-FV-lier-tout-2018-08-20.wav	86.76	🔊
Maximilien Guérin	StLeger-Dauby-0013-FV-acheter-tout-2018-08-20.wav	14.92	🔊
Maximilien Guérin	StLeger-Dauby-0014-NOM-vegetal-2018-08-20.wav	291.24	🔊
Maximilien Guérin	StLeger-Dauby-0015-NOM-nourriture-2018-08-20.wav	47.32	🔊

Les parlers du Croissant
PROJET FINANCIÉ PAR L'ANR
ACCORDEZ VOTRE VOIX !
ANR
Labex EFL
Accès privé : déconnecté

Corpus Building: Oral part

- Audiobooks Crescent: texts

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

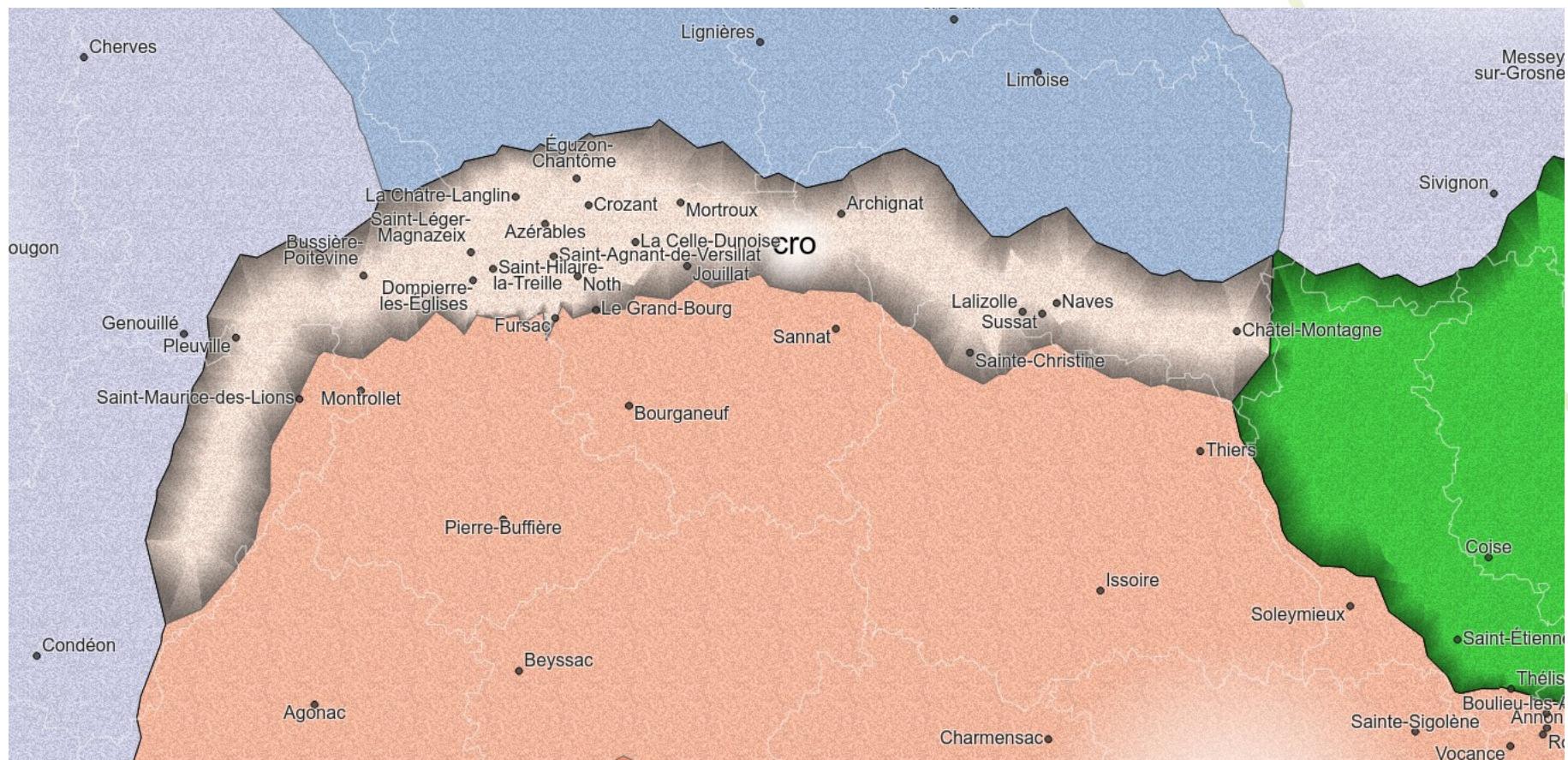
The screenshot shows a website for Bas-Marchois language resources. At the top, there are logos for ANR (Agence Nationale de la Recherche), CNRS, and EFL (Labex). On the left, there's a sidebar with book covers for 'Mes mille premiers mots en bas-marchois' and 'Contes et histoires en parler de Naves (Allier)'. The main content area has a title 'Mes mille premiers mots en bas-marchois' and author information 'Maximilien Guérin & Michel Dupeux'. Below this is a text block about the book's purpose and a reference section. A 'Lecteurs:' section mentions Jean-Michel Dauby and Monique Dauby. At the bottom, there's a table of audiobooks with columns for Author, File, Department, Commune, Size, and Listen button. The listen buttons are yellow play icons.

Auteur	Fichier	Département	Commune	Taille Mo	Ecoute
Maximilien Guérin	01_1000mots_bas-marchois_maison-p4.wav	Haute-Vienne(87)	Saint-Léger-Magnazeix	7.91	
Maximilien Guérin	02_1000mots_bas-marchois_maison-p5.wav	Haute-Vienne(87)	Saint-Léger-Magnazeix	10.85	
Maximilien Guérin	03_1000mots_bas-marchois_cuisine-p6.wav	Haute-Vienne(87)	Saint-Léger-Magnazeix	7.48	
Maximilien Guérin	04_1000mots_bas-marchois_cuisine-p7.wav	Haute-Vienne(87)	Saint-Léger-Magnazeix	10.36	
Maximilien Guérin	05_1000mots_bas-marchois_jardin-p8.wav	Haute-Vienne(87)	Saint-Léger-Magnazeix	7.29	

Corpus Building: Oral part

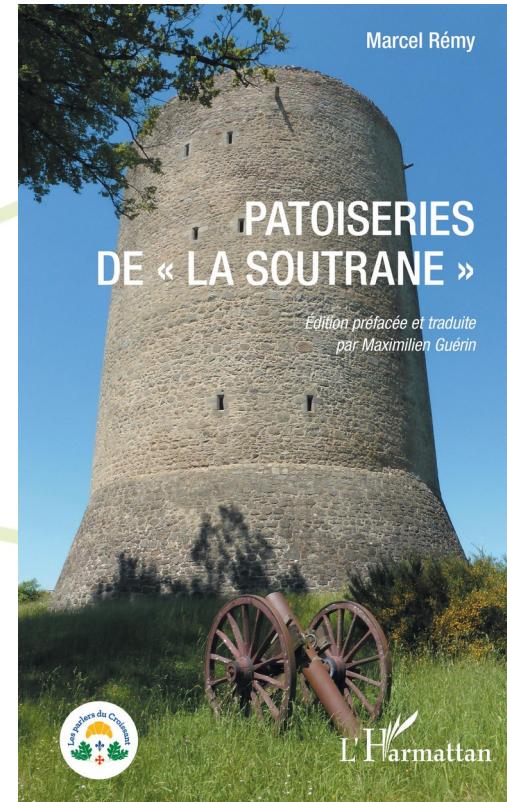
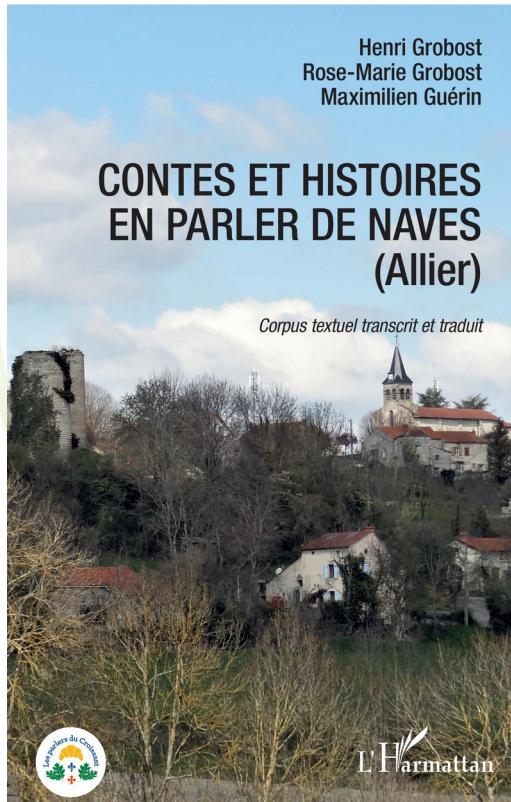
- Speaking atlas of the regional languages of France

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



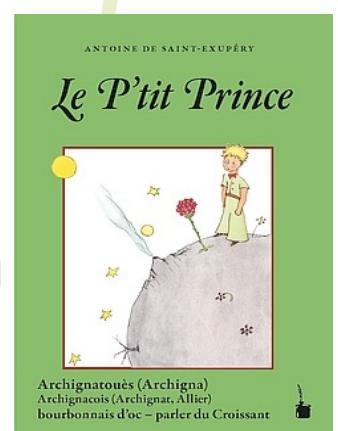
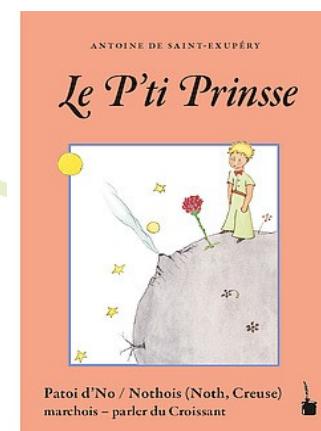
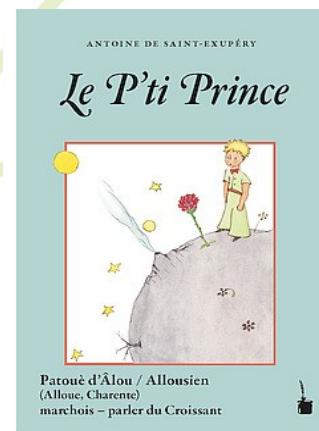
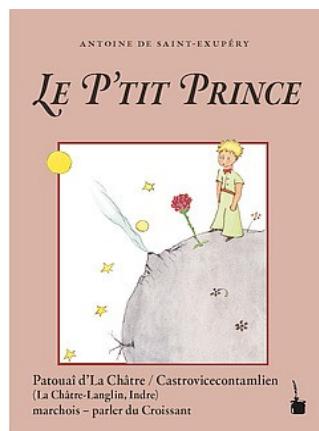
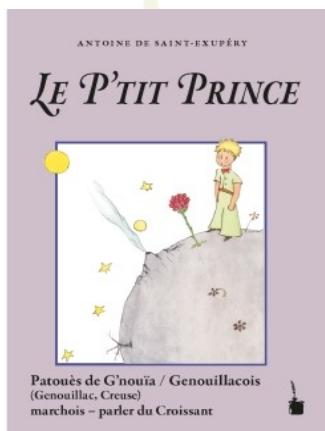
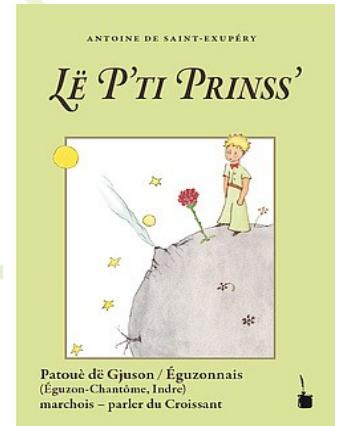
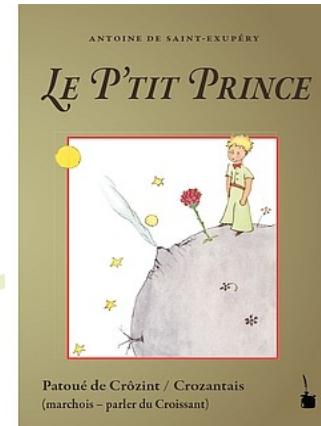
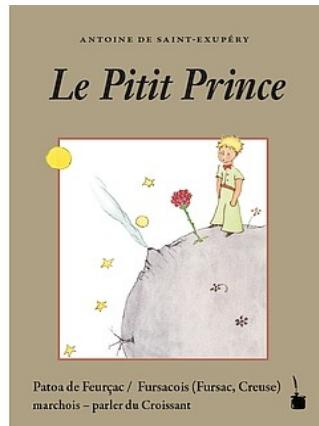
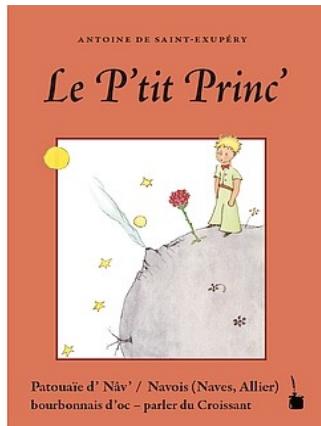
Corpus Building: Written part

- 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),
FINANCIÉS 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 choice: 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)

CORPUS CROISSANT ☰

DÉPARTEMENTS ET COMMUNES

Selectionner Tout

Désélectionner Tout

Allier (03)

Charente (16)

Cher (18)

Creuse (23)

Haute-Vienne (87)

Tout

Chateauponsac

La Croix-sur-Gartempe

Oradour-Saint-Genest

St-Leger-Magnazeix

St-Sornin-Leulac

Indre (36)

Puy-de-Dôme (63)

Vienne (86)

Le corpus Croissant contient l'ensemble des enregistrements effectués auprès de locuteurs de différentes variétés du Croissant (ou de variétés limitrophes d'oc et d'oïl), classées par département et par commune. Il est possible de rechercher un contenu particulier en tapant n'importe quel mot faisant partie du titre des fichiers sons archivés, p.ex. le nom d'un village, une catégorie grammaticale, un champ sémantique (oiseaux, poissons...) ou un verbe conjugué (être, avoir...). Sauf cas particuliers, ces fichiers sont librement accessibles.

Rechercher :

Auteur	Fichier	Taille Mo	Ecoute
Maximilien Guérin	StLeger-Dauby-0001-NOM-corps-2018-08-20.wav	62.95	🔊
Maximilien Guérin	StLeger-Dauby-0002-NOM-animal-2018-08-20.wav	293.72	🔊
Maximilien Guérin	StLeger-Dauby-0003-FV-chanter-INF-PART-2018-08-20.wav	7.23	🔊
Maximilien Guérin	StLeger-Dauby-0004-FV-chanter-IND-PRS-2018-08-20.wav	18.45	🔊
Maximilien Guérin	StLeger-Dauby-0005-FV-chanter-IND-IPRF-2018-08-20.wav	11.47	🔊
Maximilien Guérin	StLeger-Dauby-0006-FV-chanter-PASS-2018-08-20.wav	10.85	🔊
Maximilien Guérin	StLeger-Dauby-0007-FV-chanter-FUT-2018-08-20.wav	13.28	🔊
Maximilien Guérin	StLeger-Dauby-0008-FV-chanter-COND-2018-08-20.wav	11.71	🔊
Maximilien Guérin	StLeger-Dauby-0009-FV-chanter-SUBJ-PRS-2018-08-20.wav	6.7	🔊
Maximilien Guérin	StLeger-Dauby-0010-FV-chanter-SUBJ-IPRF-2018-08-20.wav	11.24	🔊
Maximilien Guérin	StLeger-Dauby-0011-FV-chanter-IMP-2018-08-20.wav	4.39	🔊
Maximilien Guérin	StLeger-Dauby-0012-FV-lier-tout-2018-08-20.wav	86.76	🔊
Maximilien Guérin	StLeger-Dauby-0013-FV-acheter-tout-2018-08-20.wav	14.92	🔊
Maximilien Guérin	StLeger-Dauby-0014-NOM-vegetal-2018-08-20.wav	291.24	🔊
Maximilien Guérin	StLeger-Dauby-0015-NOM-nourriture-2018-08-20.wav	47.32	🔊

Les parlers du Croissant
ANR
Labex EFL
Accès privé : déconnecté

Exports and data management

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

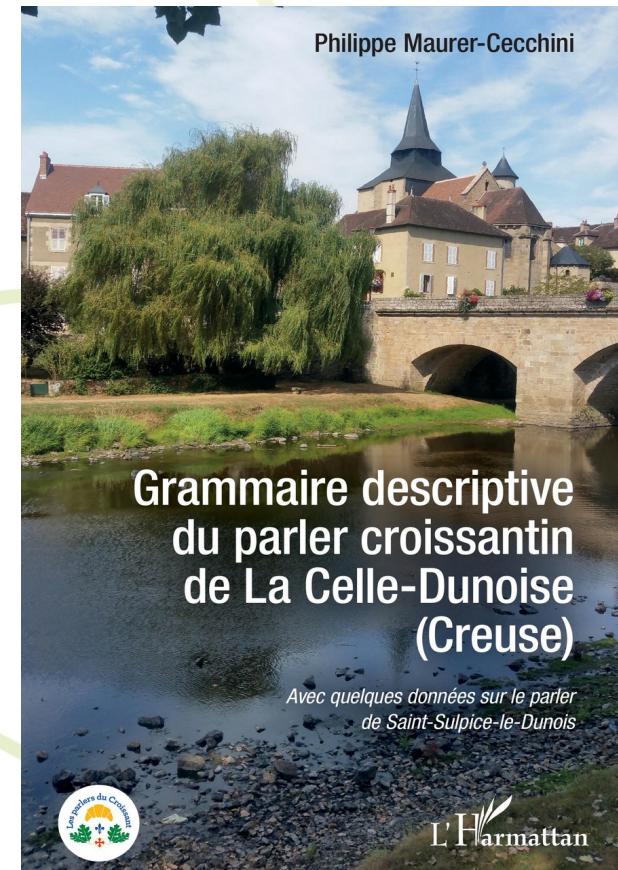
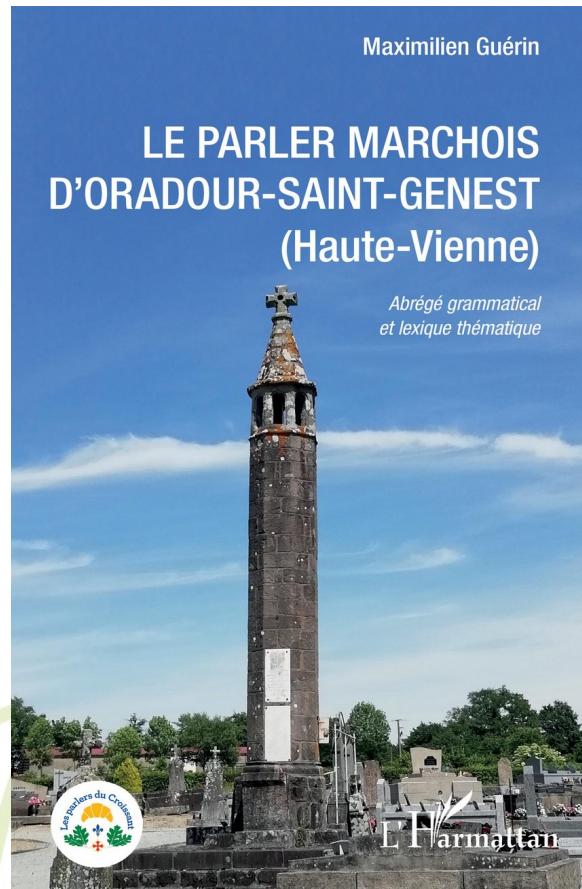
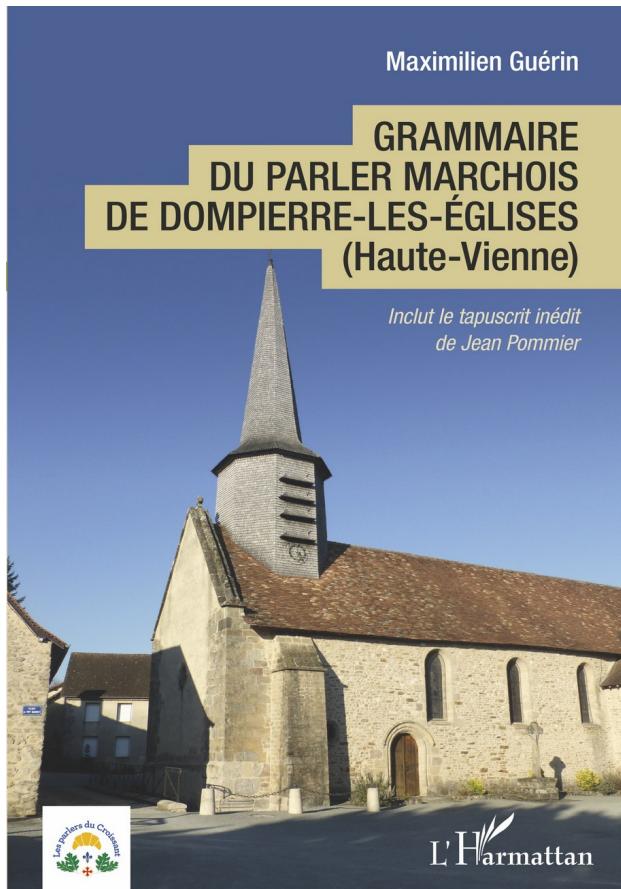
<https://parlersducroissant.huma-num.fr/conjugaison/>

The screenshot shows the 'Croissant Conjugation' website. At the top, there are logos for the Académie des sciences de l'Institut de France, ANR, and CNRS. Below the header, there's a sidebar titled 'DEMIT YOUR CHOICE' with dropdown menus for 'Variety' (Allier, Naves, Creuse, Crozant, Haute-Vienne, Dompierre-les-Eglises), 'Verb in English' (be, be able, be necessary, be worth, believe, brood, buy, come, cover, do, drink, go, have, have to / owe, hold, know [be acquainted], know [be aware], laugh), and a date range from 1800 to 1900. The main content area is titled 'Croissant Conjugation' and states: 'The classification was developed by Tourtoulon & Bringuier (1876), ranging from A1 (varieties closest to ocl) to B2 (varieties closest to oil).'. Below this is a table titled 'Results: 15' with columns: Variety, Verb in English, Verb in Occitan, Tense/Mood, Person, Form, Alternate form, and Classification. The table lists 15 entries for various verb forms across different varieties.

Variety	Verb in English	Verb in Occitan	Tense/Mood	Person	Form	Alternate form	Classification
Crozant(Creuse)	be	ésser	Imperfect indicative	2SG	't ḕr		A3
Crozant(Creuse)	have to / owe	deure	Imperfect indicative	2SG	ti 'dvɔv		A3
Crozant(Creuse)	sing	cantar	Imperfect indicative	2SG	ti jū'tov		A3
Crozant(Creuse)	take	prene	Imperfect indicative	2SG	ti prə'nɔv		A3
Crozant(Creuse)	whiten	emblanquir	Imperfect indicative	2SG	ti bjäʃ'i'sov		A3
Dompierre-les-Eglises(Haute-Vienne)	be	ésser	Imperfect indicative	2SG	t εrə		A1
Dompierre-les-Eglises(Haute-Vienne)	have to / owe	deure	Imperfect indicative	2SG	tə dəvəvə		A1
Dompierre-les-Eglises(Haute-Vienne)	sing	cantar	Imperfect indicative	2SG	tə fātəvə		A1
Dompierre-les-Eglises(Haute-Vienne)	take	prene	Imperfect indicative	2SG	tə prənəvə		A1
Dompierre-les-Eglises(Haute-Vienne)	whiten	emblanquir	Imperfect indicative	2SG	tə blāʃi'səvə		A1
Naves(Allier)	be	ésser	Imperfect indicative	2SG	t ejə		
Naves(Allier)	have to / owe	deure	Imperfect indicative	2SG	tə dyvja		
Naves(Allier)	sing	cantar	Imperfect indicative	2SG	tə jētʃja		
Naves(Allier)	take	prene	Imperfect indicative	2SG	tə poernjə		
Naves(Allier)	whiten	emblanquir	Imperfect indicative	2SG	tə blēʃi'sja		

Analysis: Examples of the results

- Grammatical descriptions

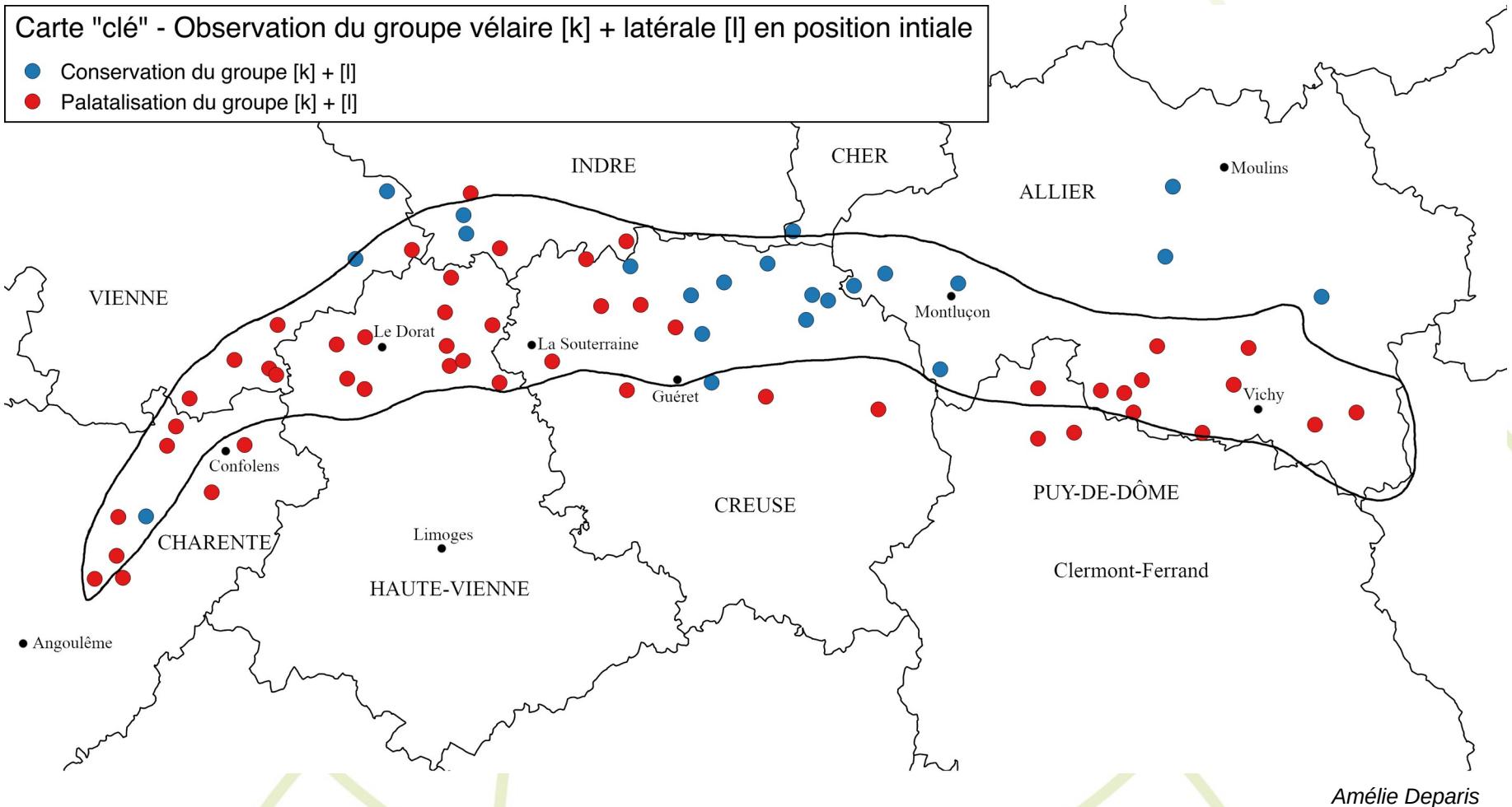


Analysis: Examples of the results

- Dialectical maps – Phonological analysis

Carte "clé" - Observation du groupe vélaire [k] + latérale [l] en position intiale

- Conservation du groupe [k] + [l]
- Palatalisation du groupe [k] + [l]

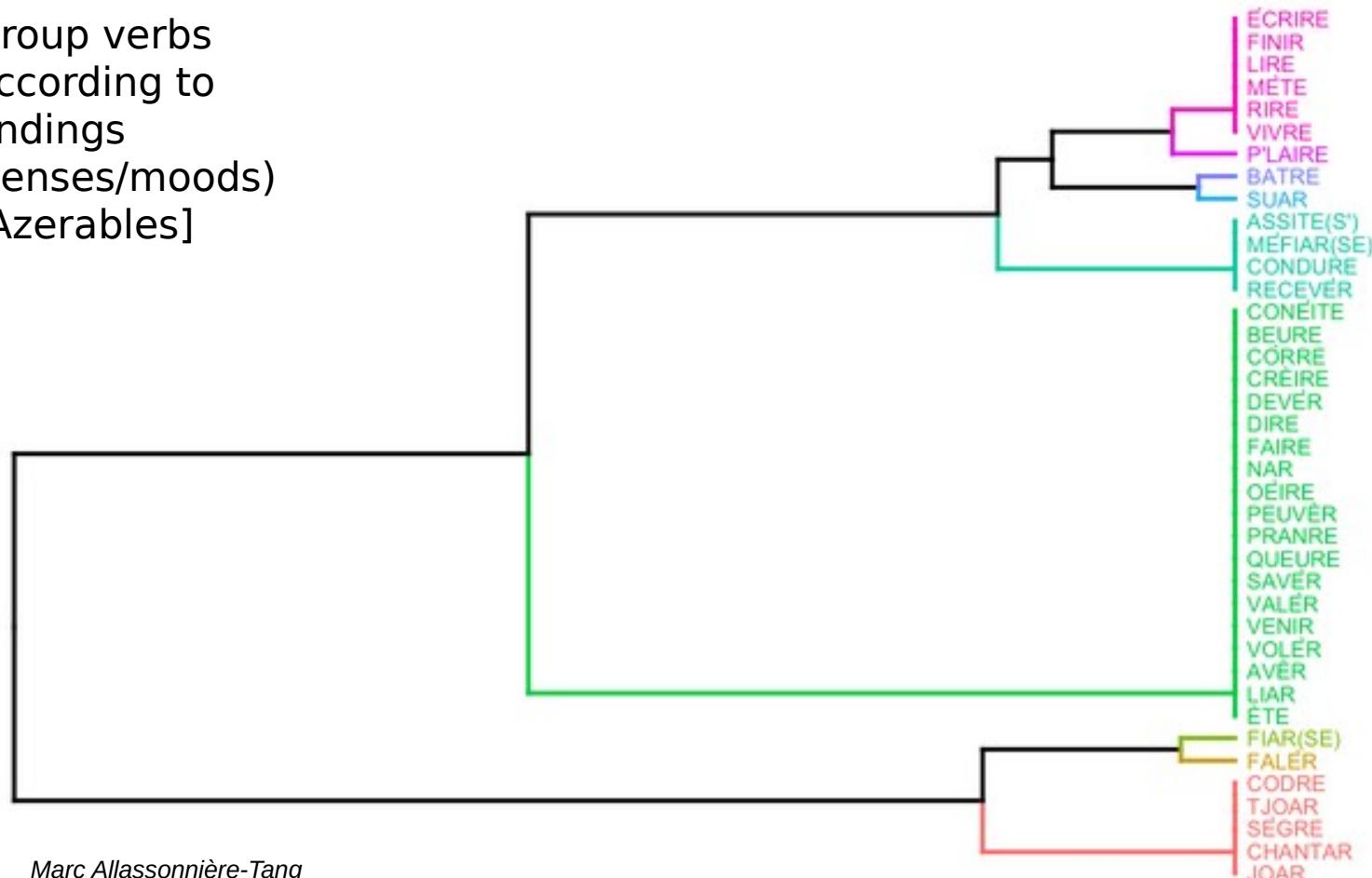


Amélie Deparis

Analysis: Examples of the results

- Hierarchical clustering - Morphological analysis

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

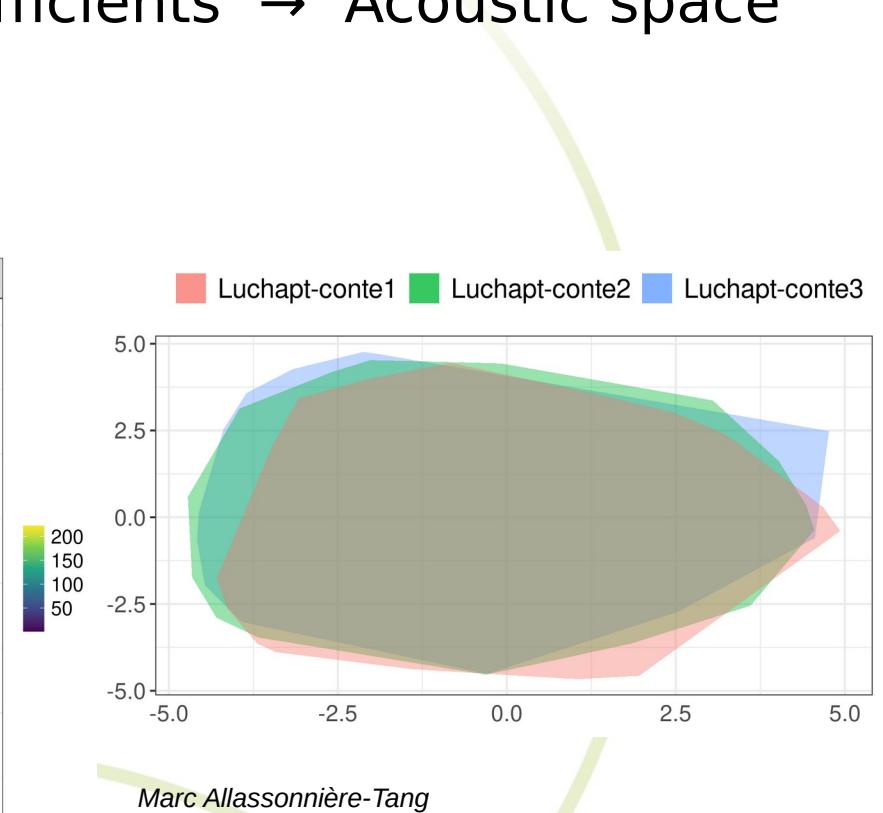
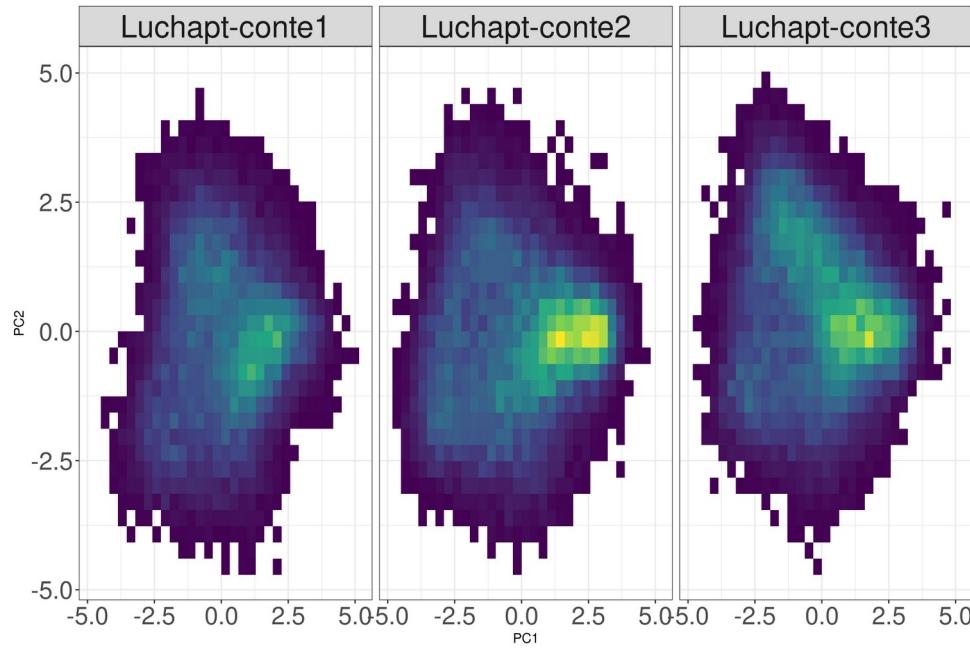


Marc Allassonnière-Tang

Analysis: Examples of the results

- Mel-frequency cepstral coefficients → Acoustic space

Phonetic comparison



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, oil 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.