OnTologies pour l'Enrichissement de l'analyse Linguistique de l'Oral

Multi-level analysis of large spoken corpora

Yaru Wu

(LISN/CNRS & LPP/CNRS)

Outline of the presentation

- My background
- 6 objectives of the OTELO project
- What I did for the OTELO project (since Oct. 2020) and what needs to be done

Teaching L2 English

↓

Real-life application of linguistic knowledge

↓

(1) Acoustic and articulatory analyses
(2) Statistical modeling

↓

(1) Teaching L2 French/etc.
(2) Qualitative & quantitative approaches for L2 perception studies

↓

(1) Large corpora phonetics
(2) ASP & ALP
(3) Statistical modeling
(4) Machine learning techniques

↓

Psycholinguistic approach (EEG recording & ERP analyses)

↓

(1) POS-tagging
(2) Advanced ML techniques

2010: Certification as a Teacher of English as a Foreign Language, Ministry of Education, China
2010: Bachelor Degree in English and Teaching English as a Foreign Language, UHS, China
2011: University Diploma in Langue et Civilisation Françaises, U. Paris Diderot, France
2012: Bachelor Degree in Applied linguistics, U. Paris Diderot, France

2014: Master Degree in LS - Phonetics and Phonology, U. Sorbonne Nouvelle, France
Master dissertation: *Utilisation de l'échographie linguale pour améliorer la réalisation du /ʁ/ français par des apprenants sinophones*
Supervisors: Cédric Gendrot (U. Sorbonne Nouvelle & LPP/CNRS) & Pierre Hallé (LPP/CNRS)

2017: Master's Degree in Teaching French as a foreign language, U. Sorbonne Nouvelle, France
Master dissertation: *Perception du /ʁ/ français par des apprenants sinophones : analyse qualitative et quantitative des expériences.*
Supervisor: Bertrand Lauret (U. Sorbonne Nouvelle)

2018: Ph.D in LS - Phonetics and Phonology, U. Sorbonne Nouvelle, France
PhD dissertation*: Étude de la réduction segmentale en français parlé à travers différents styles : apports desgrands corpus et du traitement automatique de la parole à l'étude du schwa, du /K/ et des réductions à segments multiples.*
Supervisors: Martine Adda-Decker (LPP/CNRS) & Cécile Fougeron (LPP/CNRS)

2018-2020: ATER (full-time), D. of Language Sciences, U. Paris Nanterre & Modyco/CNRS

2020-today: Post-doc at LISN/CNRS & U. Paris Saclay (Financed by DATAIA & MSH, U. Paris Saclay)

OTELO (I. Vasilescu & F. Suchanek)

- Objective of the project: modeling ambiguities in spoken language

E.g. How could language levels contribute to disambiguating the word "Paris"?

Paris /pari/ *vs* Bari /bari/ *vs* (the) bet /pari/

(city in France *vs* Italy *vs* common nouns etc.)

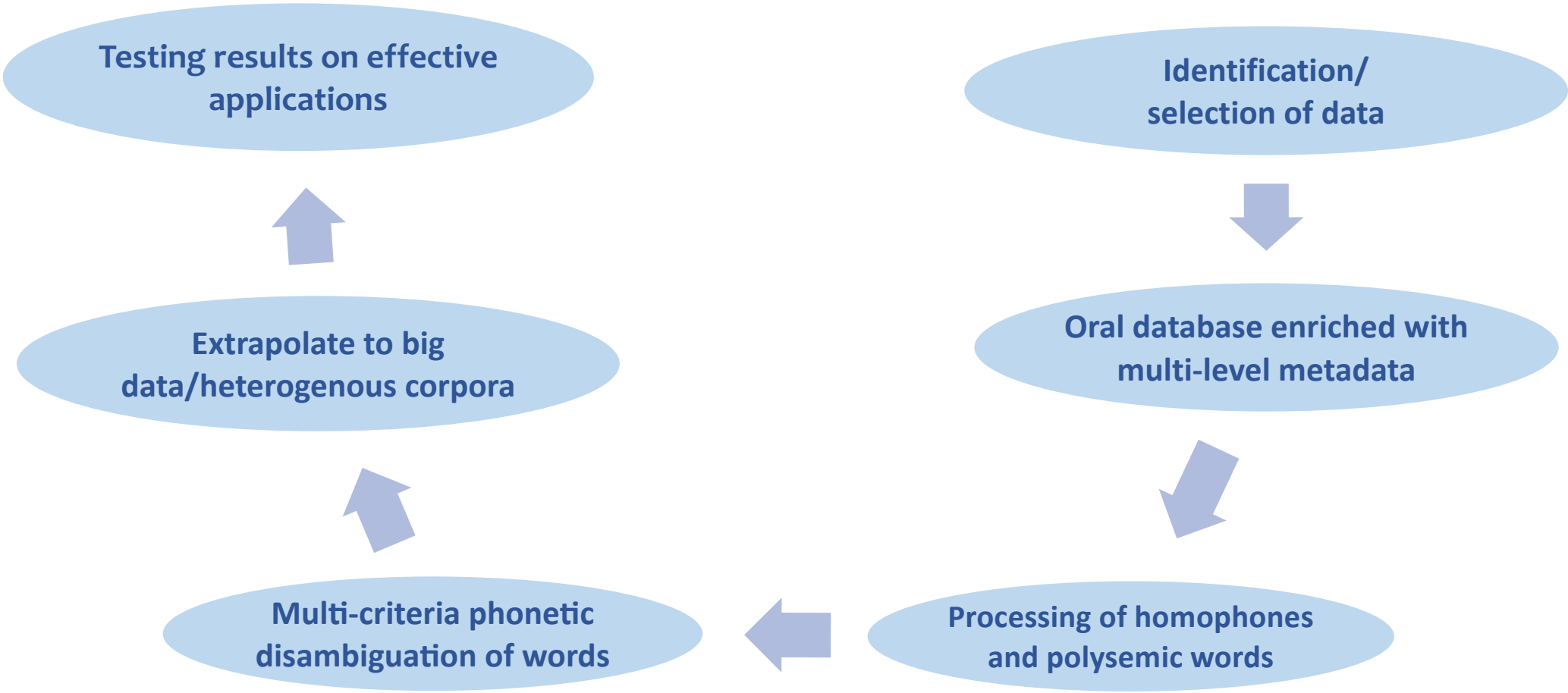
- Currently language sciences are fragmented, each community models its language level.

(e.g. semantics *vs* phonetics)

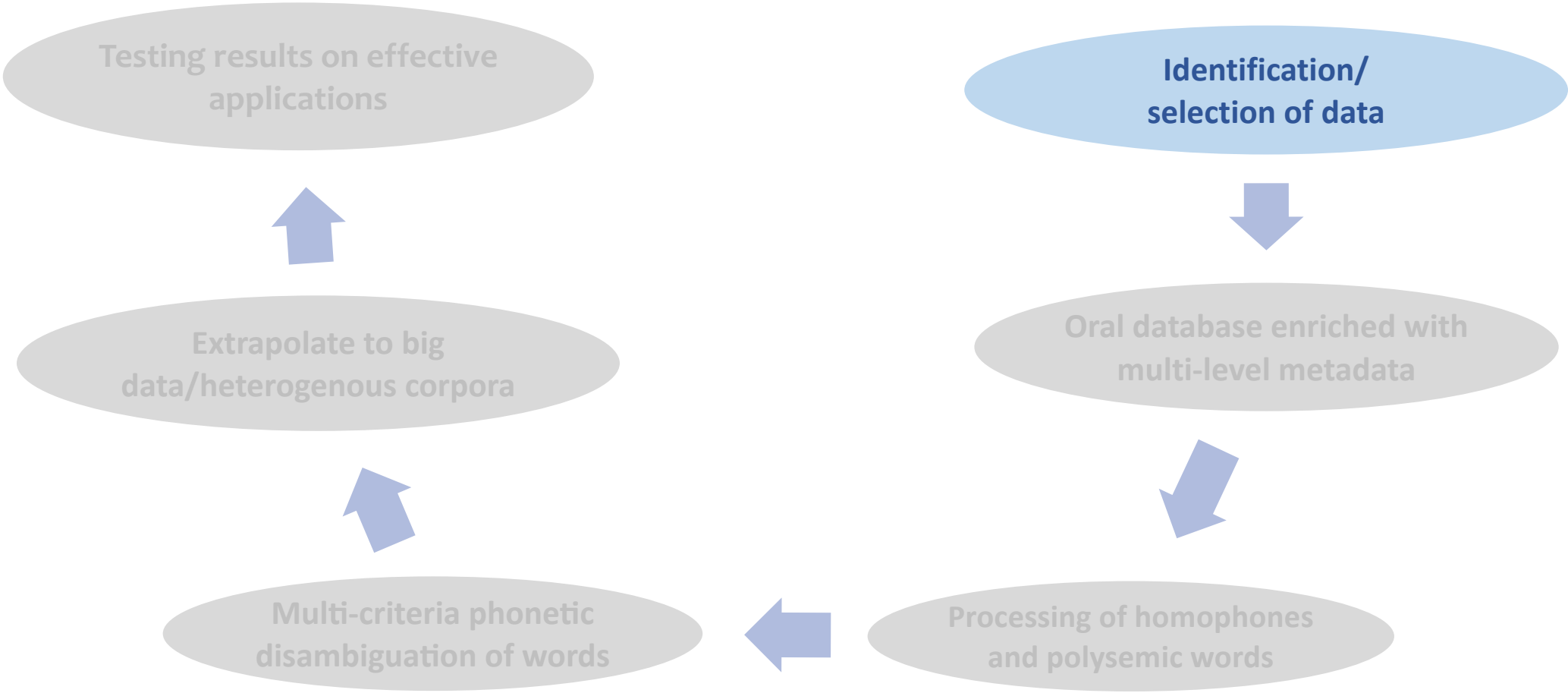
OTELO :

- Multi-level information fusion
- Resolution of ambiguity

OTELO


Identification/
selection of data
Oral database enriched with
multi-level metadata
Processing of homophones
and polysemic words
Multi-criteria phonetic
disambiguation of words
Extrapolate to big
data/heterogenous corpora
Testing results on effective
applications

OTELO


Identification/
selection of data
Oral database enriched with
multi-level metadata
Processing of homophones
and polysemic words
Multi-criteria phonetic
disambiguation of words
Extrapolate to big
data/heterogenous corpora
Testing results on effective
applications

Identification/selection of data

- Corpora from speech technology projects :
  - French: ESTER (Galliano et al. 2006), ETAPE (Gravier et al. 2012), NCCFR (Torreira et al. 2010), QUAERO
  - English: QUAERO
- Manual transcription
- Corpus manually transcribed and annotated (UCLouvain) : corpus LOCAS

Steps before multi-level annotations:

- Check if they benefited from annotations in metadata

  => yes, partially (named entities in ESTER at LISN)
- Draw a typology of speaking styles

  => Read / broadcast news / public conversations-debates / casual speech between friends
- Identify proper names
- Identify “homophony pronunciation dictionaries”

  => one phonetic transcription ⇔ corresponding orthographic transcription of homophones
- Restructure the transcription of the corpora (=> format Stanford U.)

✓ Done
❍ Ongoing
☛ To-do list

Identification/selection of data

- Corpora from speech technology projects :
  - French: ESTER (Galliano et al. 2006), ETAPE (Gravier et al. 2012), NCCFR (Torreira et al. 2010), QUAERO
  - English: QUAERO
- Manual transcription
- Corpus manually transcribed and annotated (UCLouvain) : corpus LOCAS

Steps before multi-level annotations:

✓ Check if they benefited from annotations in metadata
=> yes, partially (named entities in ESTER at LISN)

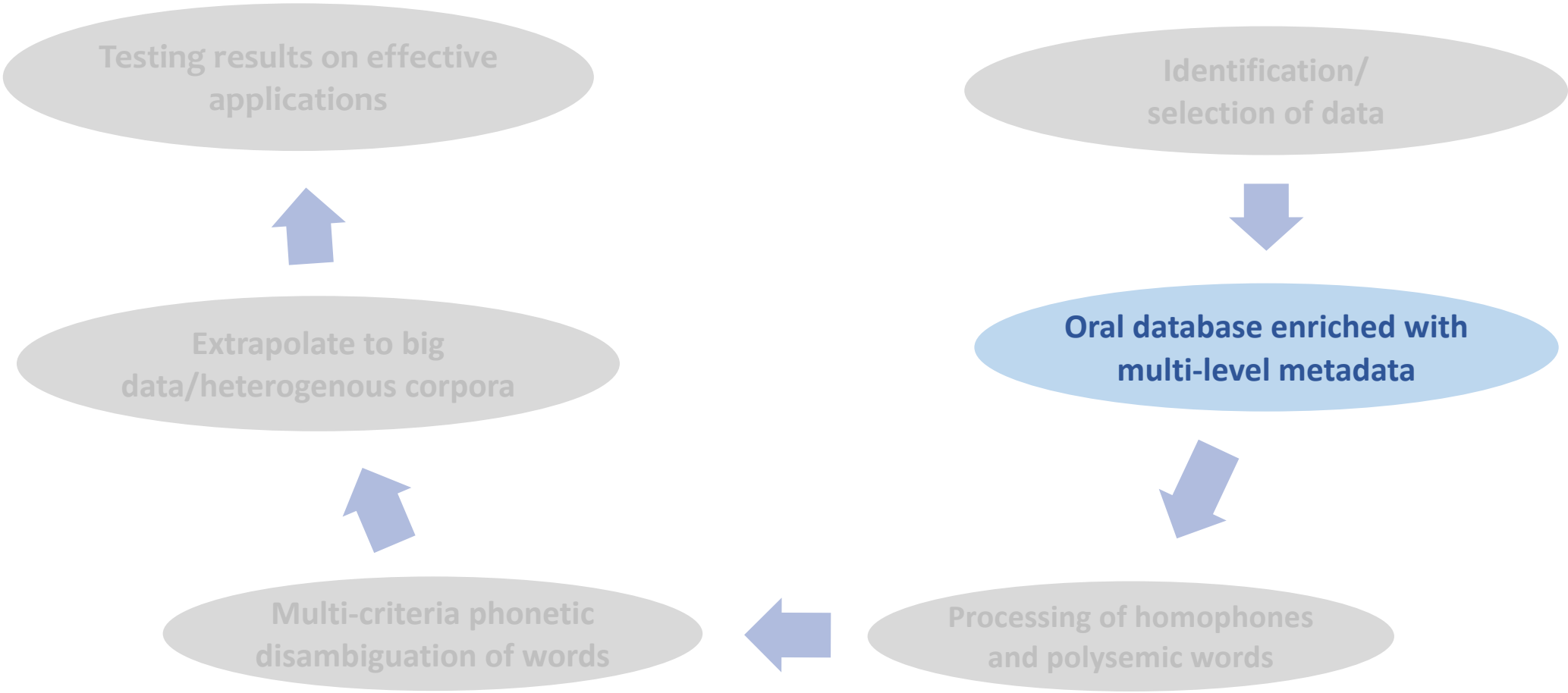
✓ Draw a typology of speaking styles
=> Read / broadcast news / public conversations-debates / casual speech between friends

✓ Identify proper names

✓ Identify "homophony pronunciation dictionaries"
=> one phonetic transcription ⇔ corresponding orthographic transcription of homophones

✓ Restructure the transcription of the corpora (=> format Stanford U.)

OTELO


Identification/
selection of data
Oral database enriched with
multi-level metadata
Processing of homophones
and polysemic words
Multi-criteria phonetic
disambiguation of words
Extrapolate to big
data/heterogenous corpora
Testing results on effective
applications

Oral database enriched with multi-level metadata

- POS tagging using STANZA (Stanford University)
  - Extraction of POS for each entry (word-token) of the spoken corpora
  - Mapping added POS to the phone level annotations
- YAGO-based socio-linguistic information extraction

  (developed by Fabian Suchanek): age, speaker origin, education, etc.
  - Tests on French data
  - Manual evaluations of output
  - Identification of data related bias (unknown speaker with famous names)

    => call for further information extraction methods (NoRDF)
- NoRDF (12/2020 -)
  - NoRDF: on-going project at Télécom Paris, supervised by F. Suchanek & C. Clavel
  - Modeling and extraction of complex information from text in natural language

✓ Done
❍ Ongoing
☛ To-do list

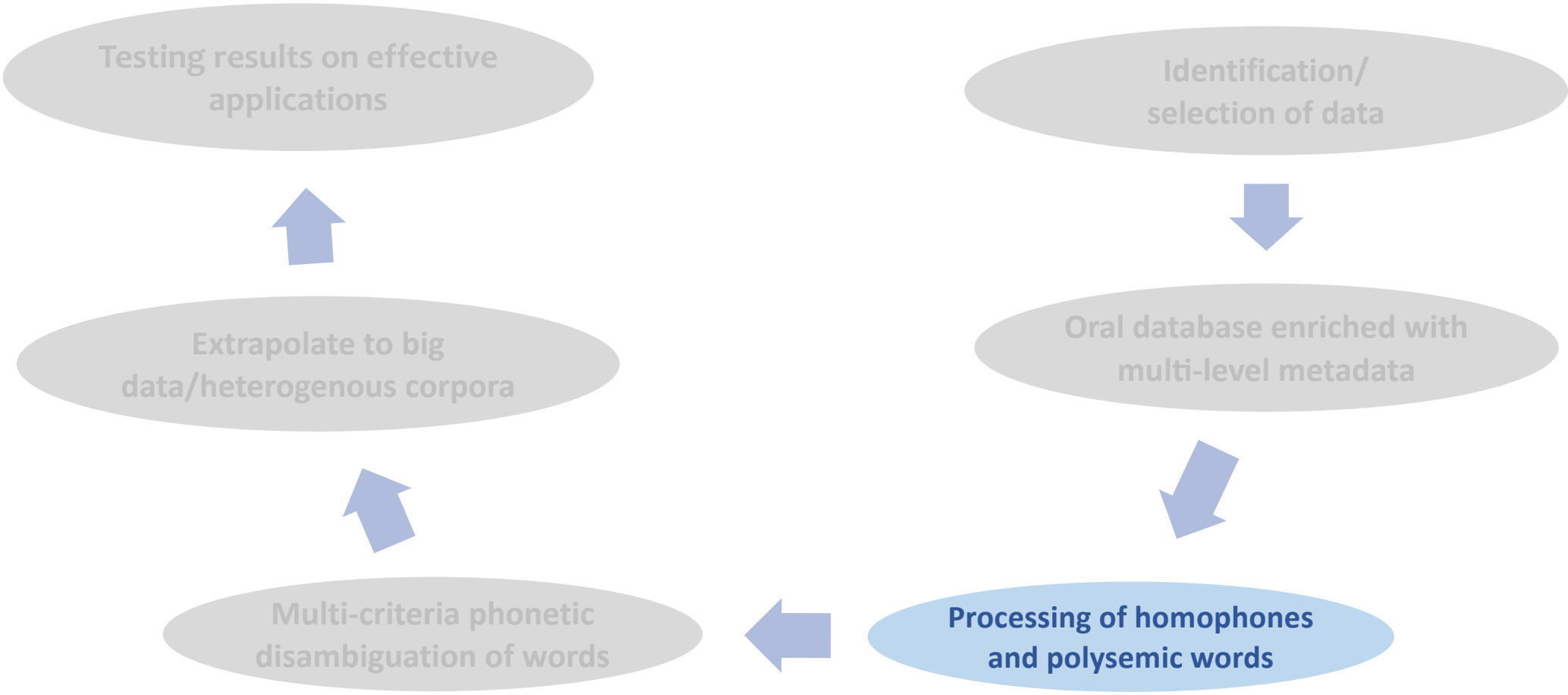
Oral database enriched with multi-level metadata

- POS tagging using STANZA (Stanford University)
  - ✓ Extraction of POS for each entry (word-token) of the spoken corpora
  - ✓ Mapping added POS to the phone level annotations
- YAGO-based socio-linguistic information extraction

  (developped by Fabian Suchanek): age, speaker origin, education, etc.
  - ❍ Tests on French data
  - ❍ Manual evaluations of output
  - ❍ Identification of date related bias (infamous speaker with famous names)

    => call for further information extraction methods (NoRDF)
- NoRDF (12/2020 -)
  - ❍ NoRDF: on-going projet at Télécom Paris, supervised by F. Suchanek & C. Clavel
  - ❍ Modeling and extraction of complex information from text in natural language

OTELO


Testing results on effective applications
Identification/ selection of data
Extrapolate to big data/heterogenous corpora
Oral database enriched with multi-level metadata
Multi-criteria phonetic disambiguation of words
Processing of homophones and polysemic words

Processing of homophones and polysemic words

Discourse markers in large scale corpora

- **Why focus on homophones and polysemic words?**

⇒Homophones and polysemic words are known sources of ambiguity for automatic systems in connected speech

(e.g. ASR error : Barack Obama/baraque aux Bahamas – *problem of "Out of vocabulary/OOV"*)

⇒We want to draw a typology of features that can help disambiguating homophones and polysemic words

- Duration
- Prosodic position
- formant frequencies in the vowels
- Harmonic to Noise Ratio (HNR)
- Etc.

- **Discourse markers (e.g. alors: adv. *vs* discours markers) ?**

=> LOCAS: 4 hours of manually annotated data (DM *vs* non-DM) => baseline for large corpora

Processing of homophones and polysemic words

Discourse markers in large scale corpora

- Methods
  - Automatic classification (decision tree)
  - Automatic extraction of cues
  - Statistical modeling with LMM/GLMM
  - Manual verification of sub corpus

(methodology between phonetic-oriented approaches : manual check and statistical modeling)

✓ Done
❍ Ongoing
☛ To-do list

Processing of homophones and polysemic words

Discourse markers in large scale corpora

- Methods
  - ✓ Automatic classification (decision tree)
  - ❍ Automatic extraction of cues
  - ❍ Statistical modeling with LMM/GLMM
  - ☛ Manual verification of sub corpus

(methodology between phonetic-oriented approaches : manual check and statistical modeling)

[Automatic classification of discourse markers]

Methods

Step 1: Decision trees applied to the LOCAS corpus

- Factors: durations, POS
- Train: 70%; test 30%

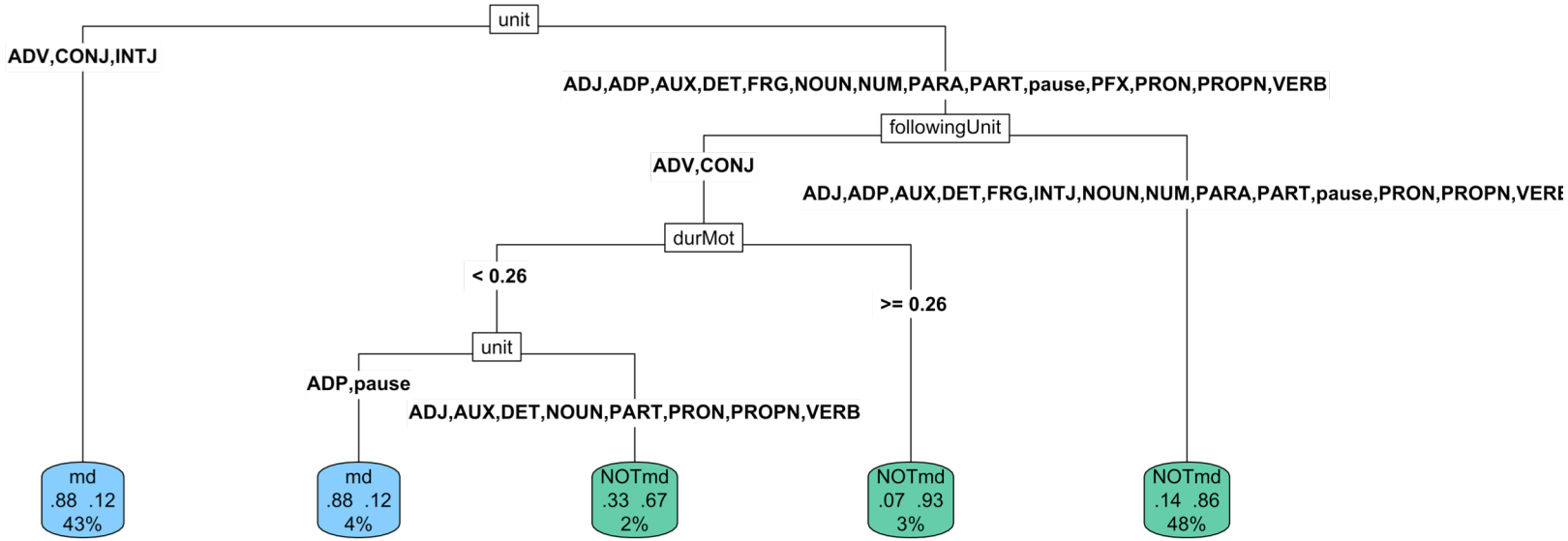
Step 2: Predicting "discourse markers" status of lexical items in 4 large scale corpora

Typology of most predictive features

[Automatic classification of discourse markers]

Features predicting discours markers in LOCAS (Step 1)

- DM => ADJ, CONJ, INTJ, PREP



[Automatic classification of discourse markers]

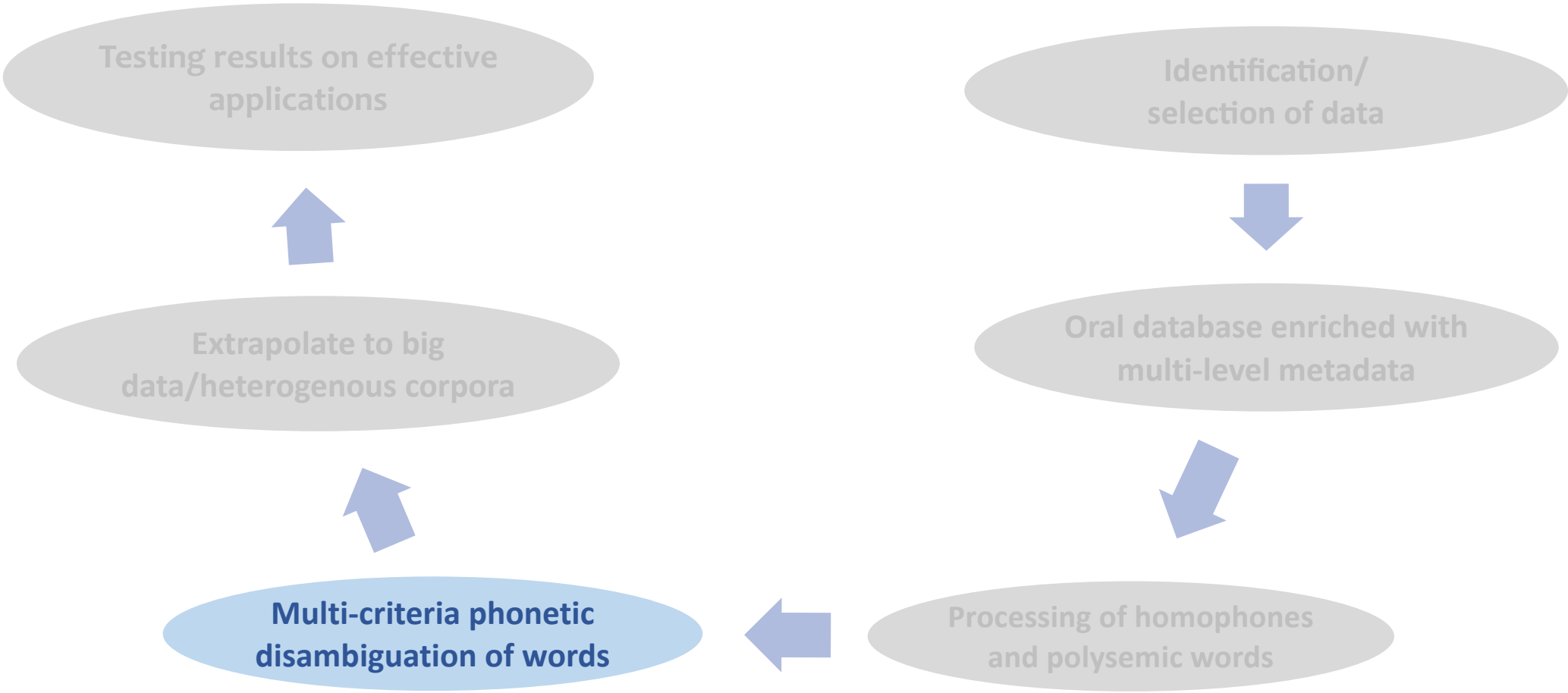
Ongoing analyses (Extension of Step 1)

- Extending the features to acoustic cues and evaluation of the algorithm prediction
- Description and typology of acoustic properties of discourse markers

Step 2:

Discourse marker identification and characterization in large scale corpora using the described method

OTELO


Identification/
selection of data
Oral database enriched with
multi-level metadata
Processing of homophones
and polysemic words
Multi-criteria phonetic
disambiguation of words
Extrapolate to big
data/heterogenous corpora
Testing results on effective
applications

Multi-criteria phonetic disambiguation of words

- **Morpho-syntactic and discursive approaches**

  =>build classes and groups of words according to
    - POS
    - discourse status
    - Acoustic cues
    - socio-linguistic factors
- **Phonetic approach**

  => build a typology of variation patterns => non-canonical realization of words
    - Reduction
    - Adjacency phenomena (e.g. assimilation/dissimilation etc)

Purposes:

- Propose methods that allow to model variation (efficiently) as function of word class and source of data (socio-linguistic factors)
- Predict "alteration" phenomena with respect to canonical realizations / grammatical class / discursive role

Multi-criteria phonetic disambiguation of words

- **Morpho-syntactic and discursive approaches**

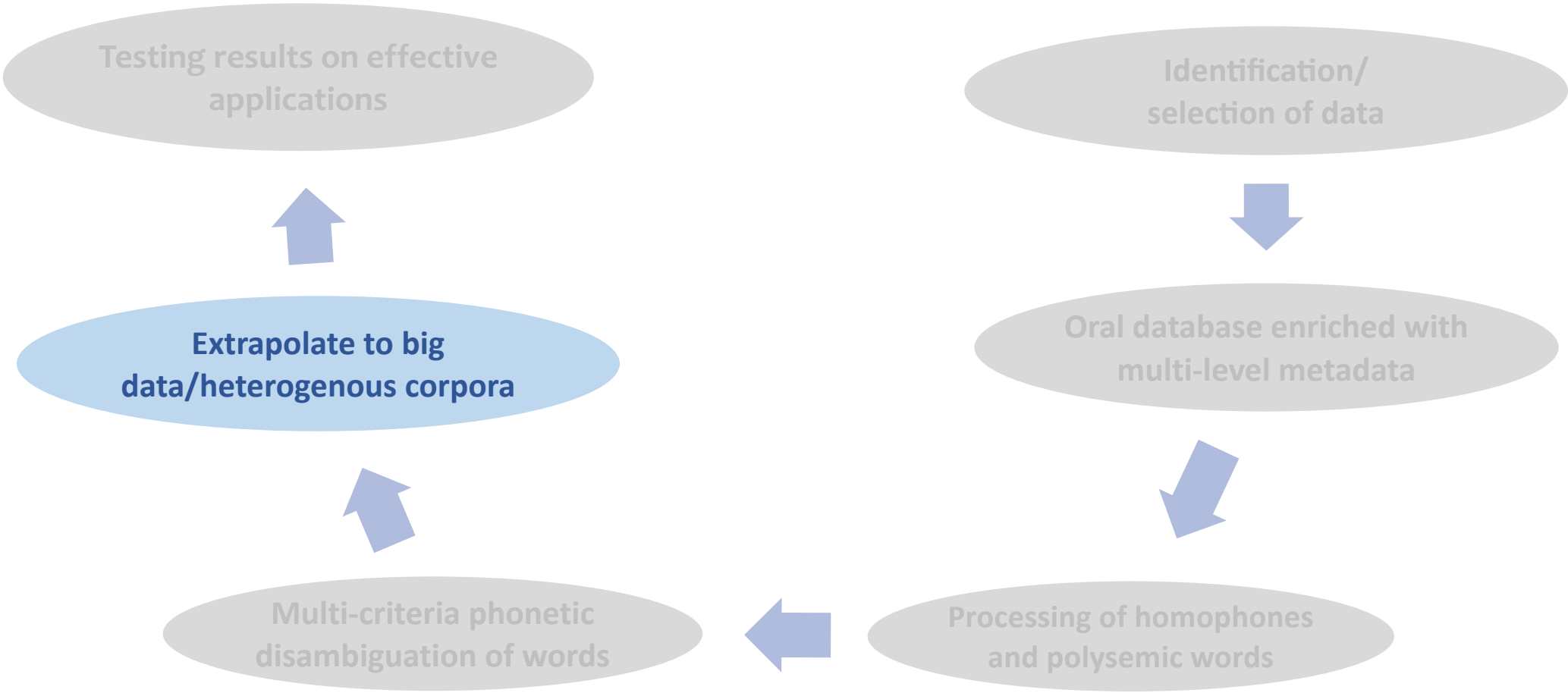
  =>build classes and groups of words according to
  - POS
  - discourse status
  - Acoustic cues
  - socio-linguistic factors
- **Phonetic approach** (cf. Mathilde Hutin's talk in a few minutes)

  => build a typology of variation patterns => non-canonical realization of words
  - Reduction
  - Adjacency phenomena (e.g. assimilation/dissimilation etc)

Purposes:

- Propose methods that allow to model variation (efficiently) as function of word class and source of data (socio-linguistic factors)
- Predict "alteration" phenomena with respect to canonical realizations / grammatical class / discursive role

OTELO


Identification/
selection of data
Oral database enriched with
multi-level metadata
Processing of homophones
and polysemic words
Multi-criteria phonetic
disambiguation of words
Extrapolate to big
data/heterogenous corpora
Testing results on effective
applications

Extrapolate to big data/heterogenous corpora

- Extrapolate features and methods from controlled/semi-controlled to uncontrolled data
  - A first step in this direction: use of 1000 hours of speech in 5 Romance languages to model non canonical pronunciation of words
  - Focus on voicing alternations: lenition (when a voiceless consonant p/t/k is pronounced voiced) and fortition (when a voiced consonant b/d/g is devoiced)

    E.g. Vasilescu I., Wu Y., Jatteau A., Adda-Decker M., & Lamel L. (2020). Alternances de voisement et processus de lénition et de fortition: une étude automatisée de grands corpus en cinq langues romanes. *Revue TAL* (Volume 61, Numéro 1).
  - Identify heterogenous data

  (e.g. training data for ASR that are not manually transcribed, data from the Internet)
  - Use automatically (instead of manually) transcribed data
  - Use methods that rely on artificial intelligence (machine learning etc.)

✓ Done
❍ Ongoing
☛ To-do list

Extrapolate to big data/heterogenous corpora

- Extrapolate features and methods from controlled/semi-controlled to uncontrolled data
  - ❍ A first step in this direction: use of 1000 hours of speech in 5 Romance languages to model non canonical pronunciation of words
    - Focus on voicing alternations: lenition (when a voiceless consonant p/t/k is pronounced voiced) and fortition (when a voiced consonant b/d/g is devoiced)

      E.g. Vasilescu I., Wu Y., Jatteau A., Adda-Decker M., & Lamel L. (2020). Alternances de voisement et processus de lénition et de fortition: une étude automatisée de grands corpus en cinq langues romanes. *Revue TAL* (Volume 61, Numéro 1).
  - ☛ Identify heterogenous data
    (e.g. training data for ASR that are not manually transcribed, data from the Internet)
  - ☛ Use automatically (instead of manually) transcribed data
  - ☛ Use methods that rely on artificial intelligence (machine learning etc.)

[Voicing alternation in 5 Romance languages]

Voicing alternation:

- Voiceless occlusives becomes voiced: /ptk/ prononcé [bdg]

  (e.g. /bak/ "ferry" => [bag] in fr; /krap/ “I crack” [krab] in rom)

- Voiced occlusives becomes voiceless: /bdg/ prononcé [ptk]

  (e.g. /bag/ "ring" => [bak] in fr; /krab/ “crab” [krap] in rom)

Objective of the project:

- Quantify the phenomena and identify the conditions that trigger voicing alternations

Research Questions:

(1) What are the language specific/language independent factors

(2) Which word position and acoustic features trigger more voicing alternations

Vasilescu I., Wu Y., Jatteau A., Adda-Decker M., & Lamel L. (2020). Alternances de voisement et processus de lénition et de fortition: une étude automatisée de grands corpus en cinq langues romanes. *Revue TAL* (Volume 61, Numéro 1).

[Voicing alternation in 5 Romance languages]

Corpora

- 5 Romance languages: (~ 1000 hours)
  - French, Italian, Spanish, Portuguese, Romanian
  - Broadcast news data (mainly standard varieties)

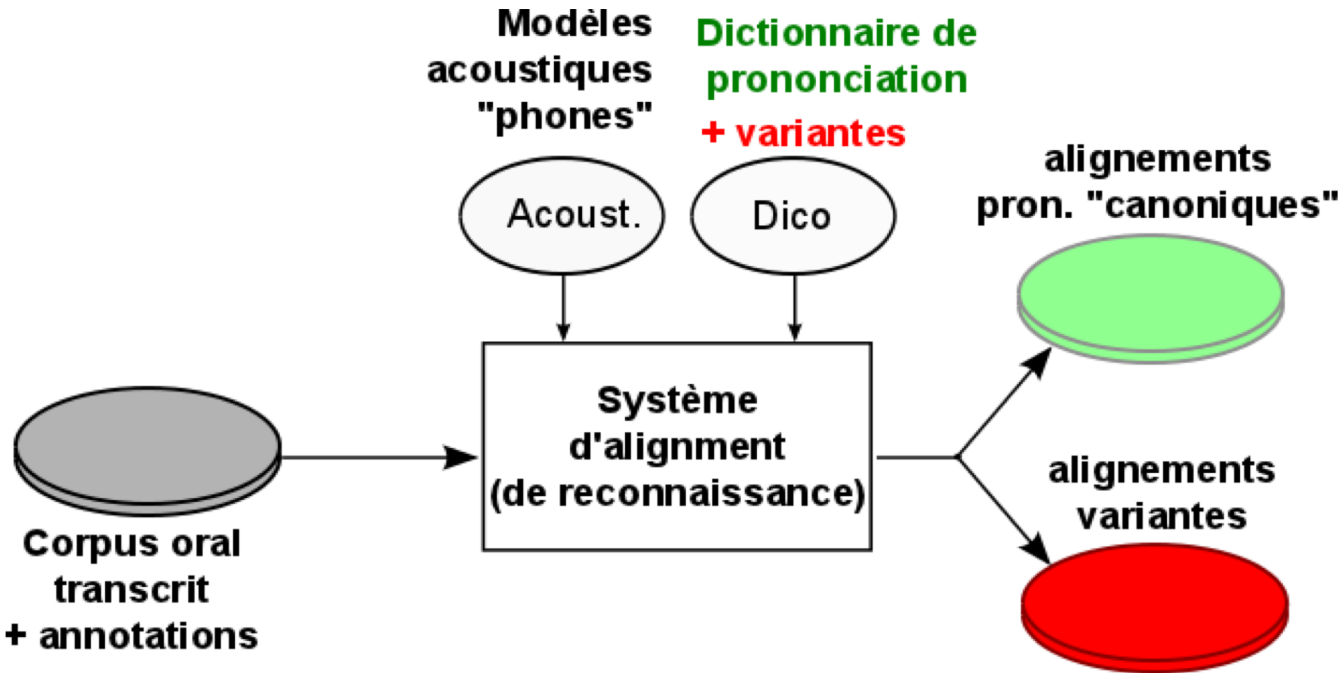
Langue	nbr. heures	mots (M)	mots uniques (k)	var/mot	var v/nv
Fr	176	2,4	55,7	2	6,8
Es	223	2,6	61,9	1	4,4
It	168	1,8	57,0	1	5,3
Po	114	1,0	40,0	1	3,7
Ro	300	3,6	48,0	1	3,7

Tableau 2. *Caractéristiques des données utilisées : langues, durée/corpus, nombre de mots (en millions M), nombre de mots uniques (en milliers k), variantes par mot dans le dictionnaire d'origine, variantes par mot permettant des alternances de voisement pour toutes les occlusives (moyennes)*

[Voicing alternation in 5 Romance languages]

Alignment

- language-specific automatic speech recognition systems (Gauvain *et al.*; 2002)
- in forced alignment mode with voicing variants (+/- voiced)



Adda-Decker et Lamel (2018)

[Voicing alternation in 5 Romance languages]

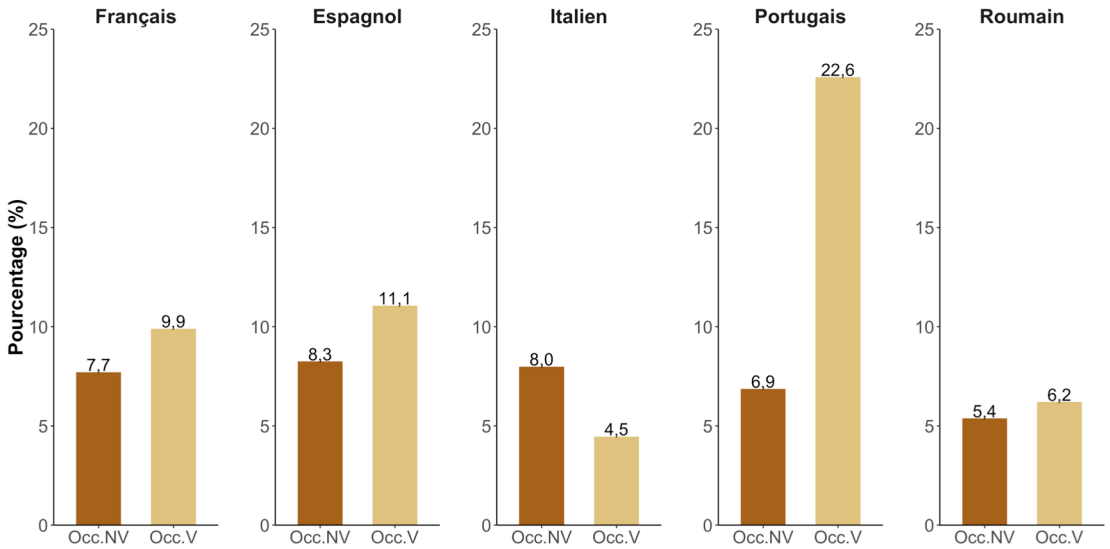
Results

All languages undergo voicing alternations

Some languages are more prone to non-canonical realizations than others.

Patterns related to position within the word and adjacency phenomena (eg. word initial -> fortition)

e.g. sac bleu => sac pleu



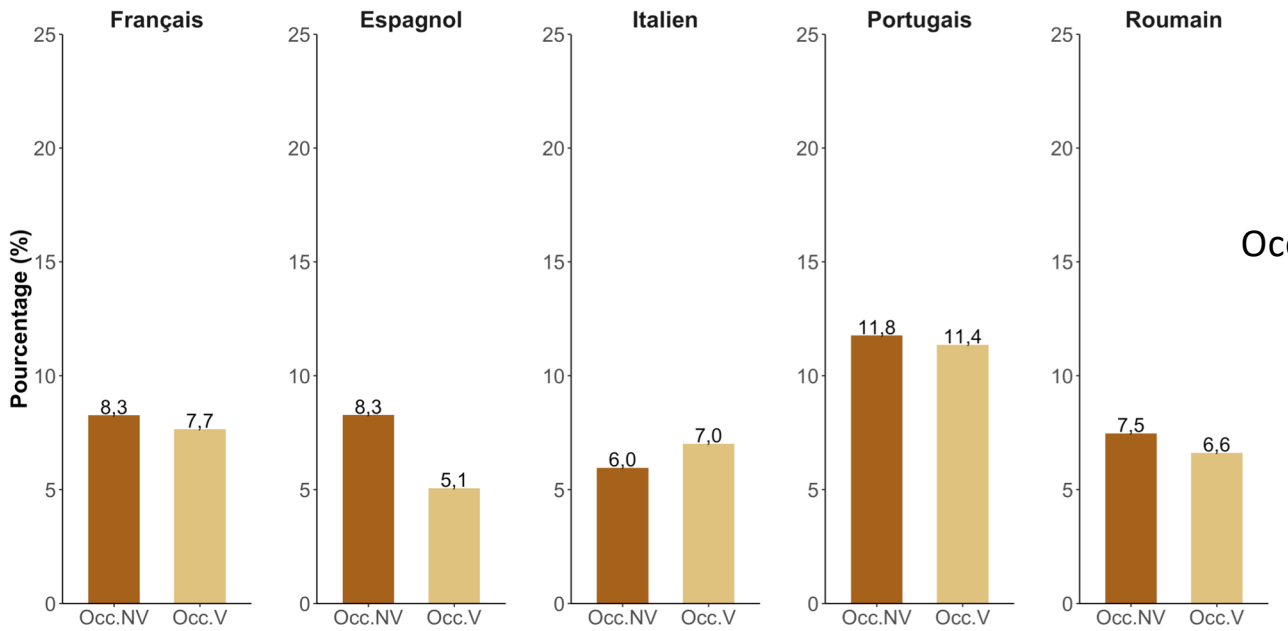
OccNV voicing < OccV devoicing ($p < 0.001$)
Except for Ita.

[Voicing alternation in 5 Romance languages]

Results:

Word internal position triggers lenition (if voiced surrounding context)

E.g. a**tt**aque : t=>[d]

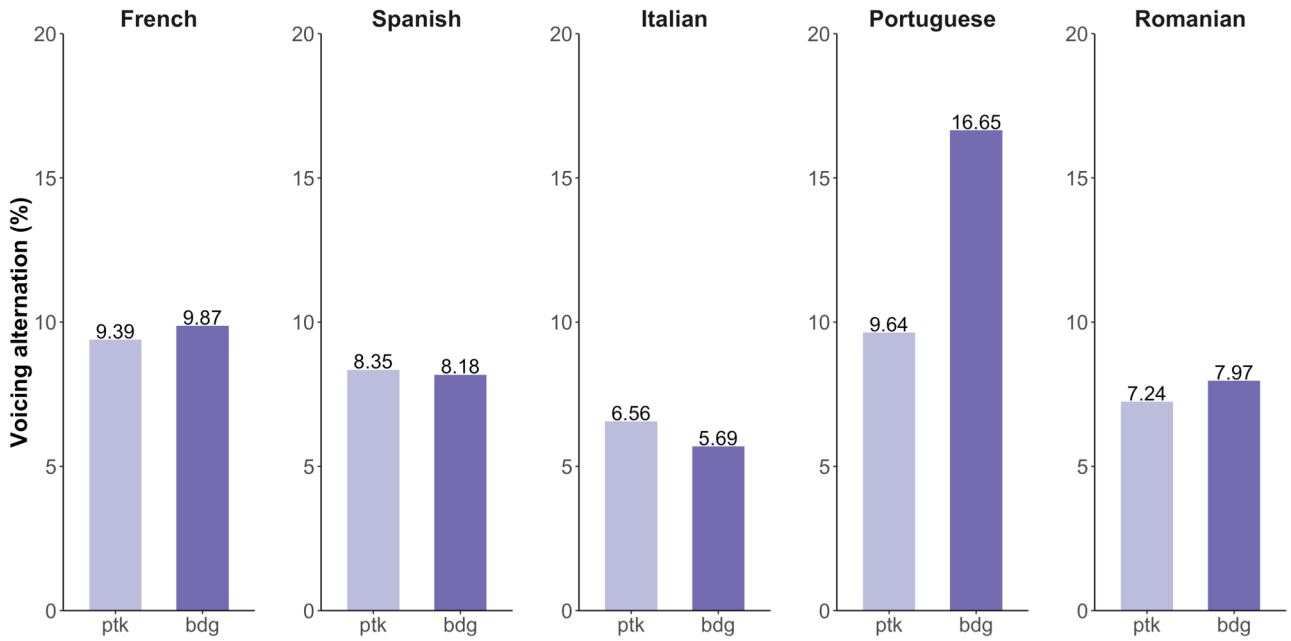


OccNV voicing > OccV devoicing ($p < 0.001$)
Except for Ita.

[Voicing alternation in 5 Romance languages]

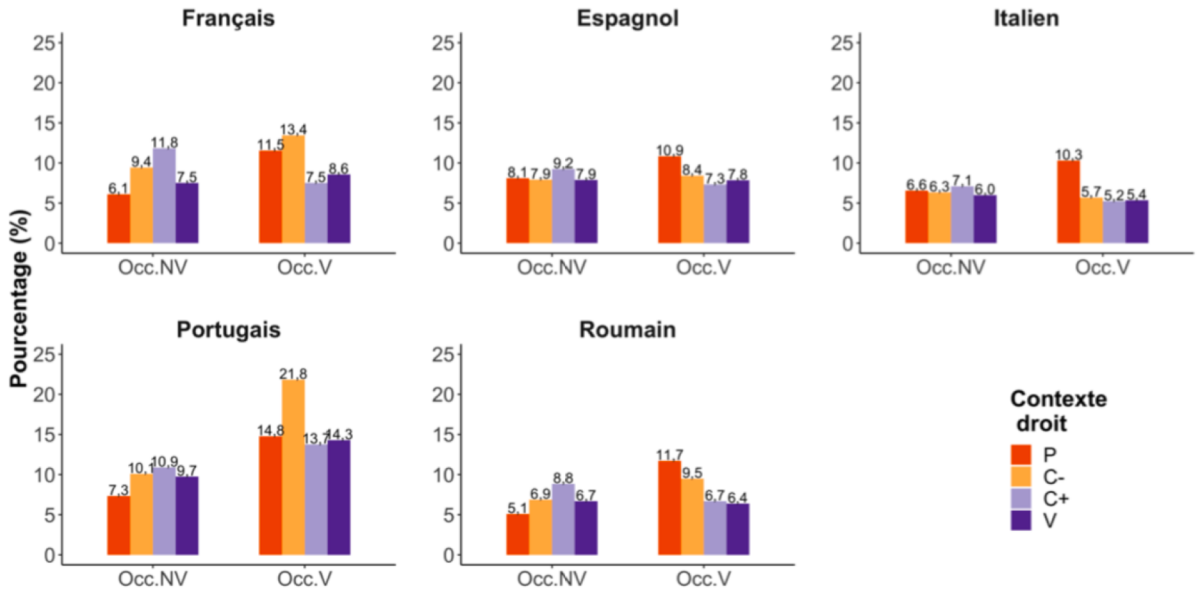
Results: In which language voicing alternations are more frequent?

- Inter language variation : Por > Fre > Ita, Spa, Rom ($p < 0.001$)



[Voicing alternation in 5 Romance languages]

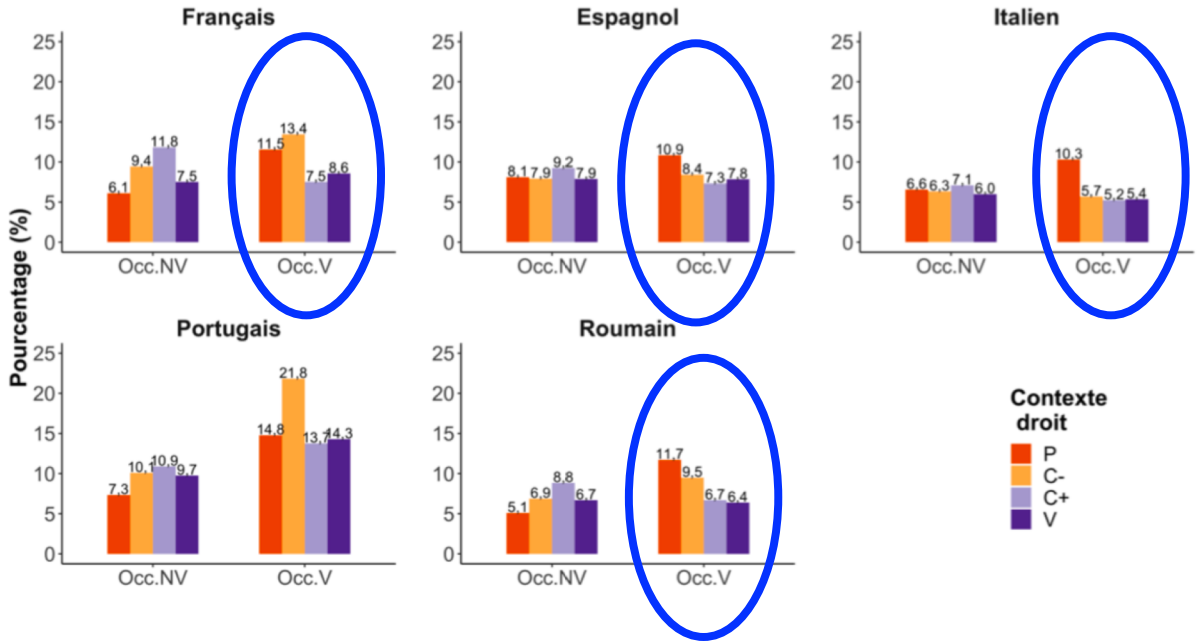
Results: Which right context triggers more voicing alternation ?



[Voicing alternation in 5 Romance languages]

Results: Which right context triggers more voicing alternation ?

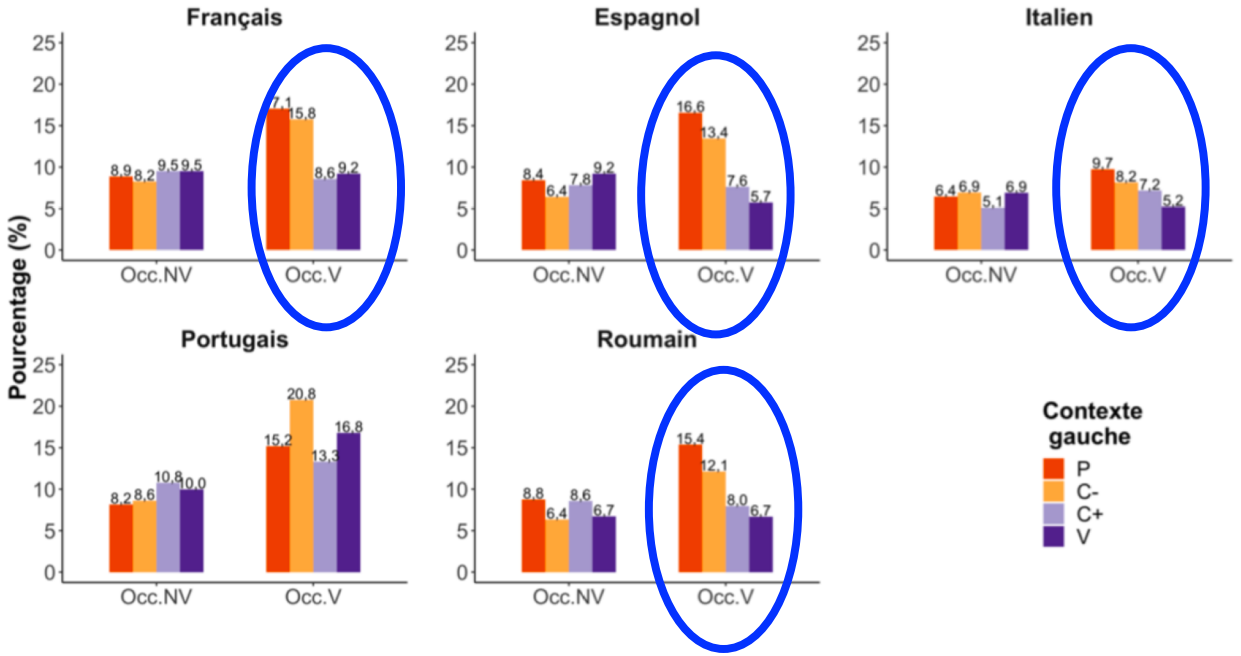
Occ.+=> Occ - : P, C- > others (except for Por.)



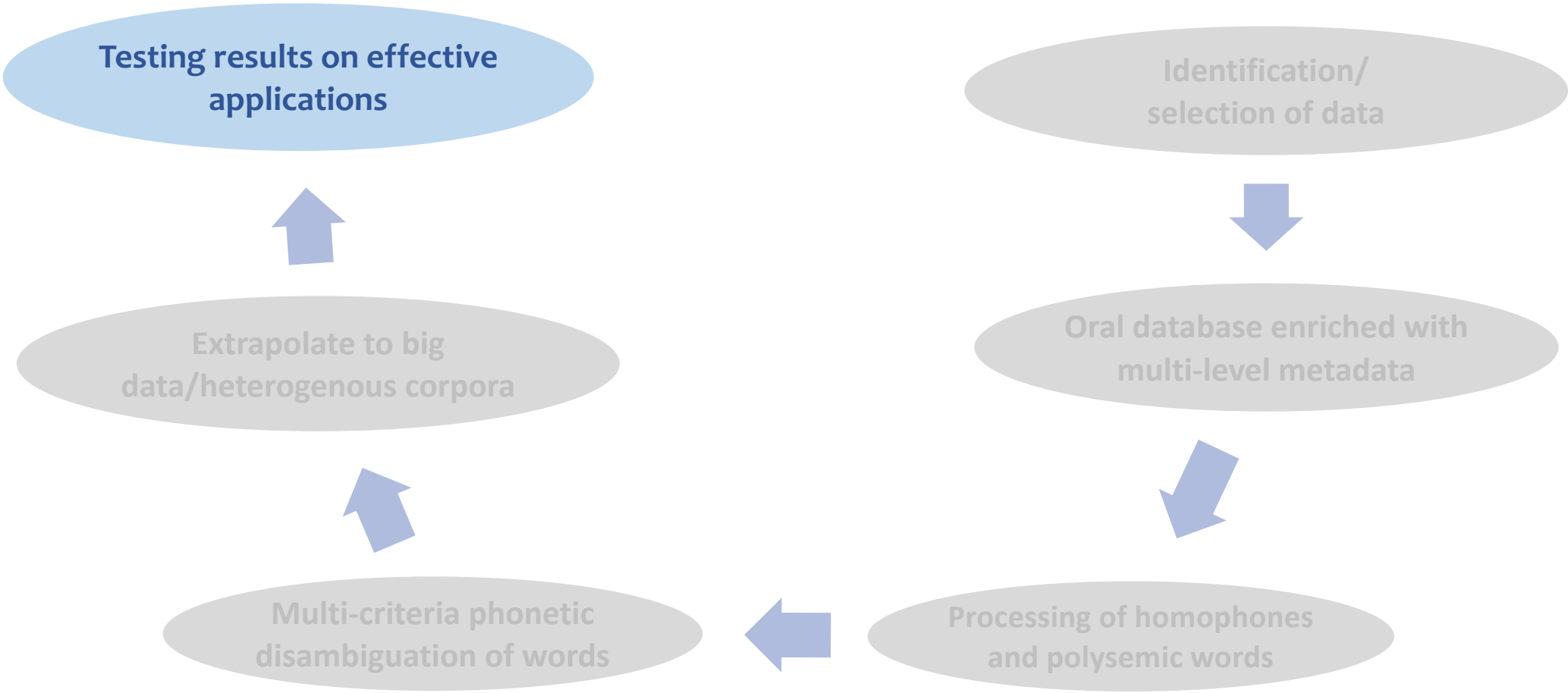
[Voicing alternation in 5 Romance languages]

Results: Which left context triggers more voicing alternation ?

+=> - : P, C- > others (except for Por.)



OTELO


Testing results on effective applications
Identification/ selection of data
Extrapolate to big data/heterogenous corpora
Oral database enriched with multi-level metadata
Multi-criteria phonetic disambiguation of words
Processing of homophones and polysemic words

Testing results on effective applications

Technology based on verbal communication (ARS, chatbots, etc.)

- Research on variation
  => improvement of Automatic Speech Recognition (ASR) dictionaries
  E.g. Vasilescu, I., Hernandez, N., Vieru, B., & Lamel, L. (2018). Exploring Temporal Reduction in Dialectal Spanish: A Large-scale Study of Lenition of Voiced Stops and Coda-s. In *INTERSPEECH* (pp. 2728-2732).

- Modeling discourse markers and disfluency
  => improvement of performance on chatbots (e.g. detection of emotion)

The last steps will be to test the outcomes on existent systems:
=> Automatic Speech Recognition (ASR) system at LISN/CNRS
=> Human-machine communication : LISN/CNRS, LTCI-Télécom Paris

Testing results on effective applications

Technology based on verbal communication (ARS, chatbots, etc.)

- Research on variation
  => improvement of ASR dictionaries
  E.g. Vasilescu, I., Hernandez, N., Vieru, B., & Lamel, L. (2018). Exploring Temporal Reduction in Dialectal Spanish: A Large-scale Study of Lenition of Voiced Stops and Coda-s. In *INTERSPEECH* (pp. 2728-2732).

- Modeling discourse markers and disfluency
  => improvement of performance on chatbots (e.g. detection of emotion)

The last steps will be to test the outcomes on existent systems:

☛ => Automatic Speech Recognition (ASR) system at LISN/CNRS

☛ => Human-machine communication : LISN/CNRS, LTCI-Télécom Paris

Publications and communications related to the OTELO project

Papers

1. **Wu Y.**, & Adda-Decker M. (in press). Réduction des segments en français spontané : apports des grands corpus et du traitement automatique de la parole. *Corpus* (N° 22).
2. **Wu Y.**, Adda-Decker M., & Lamel, L. (2020). Schwa deletion in word-initial syllables of polysyllabic words: Investigations Using Large French Speech Corpora. *Journal of Monolingual and Bilingual Speech*.
3. Vasilescu I., **Wu Y.**, Jatteau A., Adda-Decker M., & Lamel L. (2020). Alternances de voisement et processus de lénition et de fortition: une étude automatisée de grands corpus en cinq langues romanes. *Revue TAL* (Volume 61, Numéro 1).
4. **Wu, Y.**, Lamel, L., & Adda-Decker, M. (2021). Tone realization in Mandarin speech: a large corpus based study of disyllabic words. In *ISCSLP 2021*.
5. **Wu, Y.**, Adda-Decker, M., & Lamel, L. (2020). Mandarin lexical tones: a corpus-based study of word length, syllable position and prosodic structure on duration. In *INTERSPEECH 2020*.
6. **Wu, Y.**, Gendrot, C., Adda-Decker, M., & Fougeron, C. (2019). Post-consonantal Word-final /ʁ/ Realization in French: Contributions of Large Corpora. In *ICPhS 2019 (19th International Congress of Phonetic Sciences)*.
7. **Wu, Y.**, Adda-Decker, M., Fougeron, C., & Lamel, L. (2017). Schwa Realization in French: Using Automatic Speech Processing to Study Phonological and Socio-Linguistic Factors in Large Corpora. In *INTERSPEECH 2017* (pp. 3782-3786).
8. **Wu, Y.**, & Adda-Decker, M. (2017). Schwa realization in French : Investigations using automatic speech processingand large corpora. In *International Symposium on Monolingual and Bilingual Speech 2017* (pp. 294-299).
9. **Wu, Y.**, Adda-Decker, M., & Fougeron, C. (2016). Rôle des contextes lexical et post-lexical dans la réalisation du schwa: apports du traitement automatique de grands corpus. In *31èmes Journées d'Etudes sur la Parole* (No. 1, pp. 633-641).
10. Gendrot, C., Adda-Decker, M., & **Wu, Y.** (2015). Comparing Journalistic and Spontaneous Speech: Prosodic and Spectral Analysis. In *INTERSPEECH 2015* (pp. 958-962).
11. Hutin, M., Jatteau, A., **Wu Y.**, Vasilescu, I., Lamel, L., & Adda-Decker, M. (submitted). Word-final schwa in Standard French. *Journal of French Language Studies*.

Publications and communications related to the OTELO project

Communications at International conferences

1. **Wu, Y.**, Adda-Decker, M., Gendrot, C., & Lamel, L. (2019, November). Impact of post-lexical context and speechstyle on word-final /ʁ/ realization in French using large corpora and automatic speech processing. In *R-atics 6*,Paris, France.
2. **Wu, Y.**, Lamel, L. & Adda-Decker, M. (2019, September). Variation in Pluricentric Mandarin Using Large Corpus : a forced alignment-based duration and tone frequency study. In *Pluricentric Languages in Speech Technology - Satellite Workshop at Interspeech 2019*, Graz, Austria.
3. **Wu, Y.** (2019, August). Analyzing French /ʁ/ Perception in Chinese Learners Using Quantitative and QualitativeApproaches. In *New Sounds 2019*, Tokyo, Japan.
4. **Wu, Y.**, Adda-Decker, M., Fougeron, C. & Gendrot, C. (2019, June). How do French Cʁ# cluster realizations vary across speaking style?. In *Phonetics and Phonology in Europe (PaPE) 2019*, Lecce, Italy.
5. **Wu, Y.**, Gendrot, C., Adda-Decker, M., & Fougeron, C. (2018, June). Post-consonantal word-final /ʁ/ realizationin French. In *Laboratory phonology 2018*, Lisbon, Portugal.
6. Hutin, M., **Wu, Y.**, Jatteau, A., Vasilescu, I., Lamel, L., & Adda-Decker, M. (accepted). Modelling the realization of variable word-final schwa in Standard French. In *FreeVari 2021*, Freiburg, Germany.
7. Hutin, M., **Wu, Y.**, Kondo, N., Ruvoletto, S., Vasilescu, I., Lamel, L., Adda-Decker, M. (2020, November).Variabilité de la liaison facultative en français standard. In *Going Romance 2020*, Paris, France.
8. **Wu, Y.**, Hutin, M., Vasilescu, I., Lamel, L., & Adda-Decker, M. (submitted). Context, position in word and duration as predictors of voicing alternation of stops: a large-scale corpus-based study in 5 Romance Languages. In *Phonetics and Phonology in Europe (PaPE) 2021}*, Barcelona, Spain.
9. Hutin, M., **Wu, Y.**, Vasilescu, I., Lamel, L., & Adda-Decker, M. (submitted). Word-Initial Voicing Alternations inFive Romance Languages. In *Phonetics and Phonology in Europe (PaPE) 2021*, Barcelona, Spain.

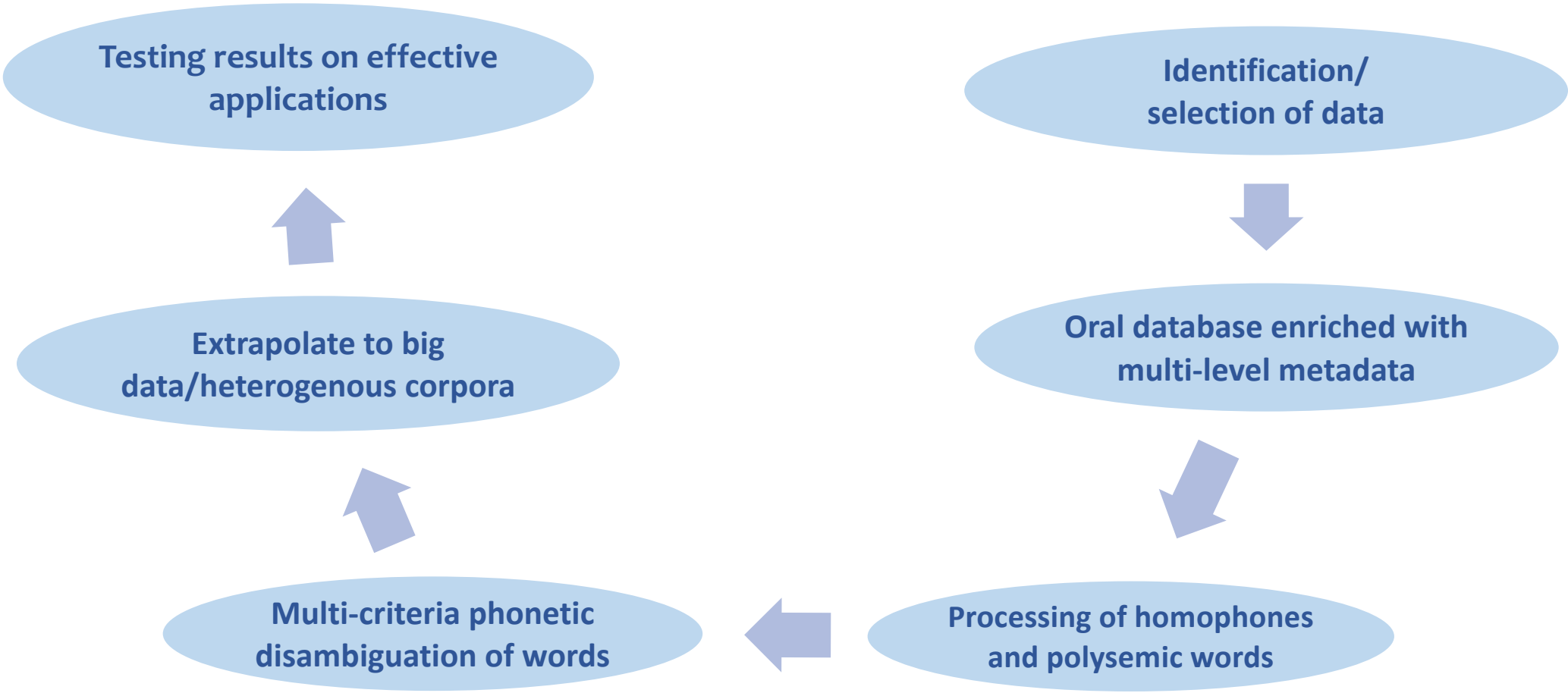
OTELO

Multi-level analysis of large spoken corpora
Communications between Humanities and Social Sciences and Digital Sciences

References

- Adda-Decker, M., & Lamel, L. (2018). Discovering speech reductions across speaking styles and languages, *Rethinking reduction - Interdisciplinary perspectives on conditions, mechanisms, and domains for phonetic variation* : 101-128.
- Galliano, S., Geoffrois, E., Gravier, G., Bonastre, J. F., Mostefa, D., & Choukri, K. (2006). Corpus description of the ESTER Evaluation Campaign for the Rich Transcription of French Broadcast News. In *LREC* (pp. 139-142).
- Gravier, G., Adda, G., Paulson, N., Carré, M., Giraudel, A., & Galibert, O. (2012). The ETAPE corpus for the evaluation of speech-based TV content processing in the French language. In *LREC-Eighth international conference on Language Resources and Evaluation*.
- Gauvain, J. L., Lamel, L., & Adda, G. (2002). The LIMSI broadcast news transcription system. *Speech communication*, *37*(1-2), 89-108.
- Lamel, L. (2012). Multilingual Speech Processing Activities in Quaero: Application to Multimedia Search in Unstructured Data. In *Baltic HLT* (pp. 1-8).
- Lamel, L., Courcinous, S., Despres, J., Gauvain, J. L., Josse, Y., Kilgour, K., ... & Woehrling, C. (2011). Speech recognition for machine translation in Quaero. In *International Workshop on Spoken Language Translation (IWSLT) 2011*.
- Torreira, F., Adda-Decker, M., & Ernestus, M. (2010). The Nijmegen corpus of casual French. *Speech Communication*, *52*(3), 201-212.
- Vasilescu, I., Rosset, S., & Adda-Decker, M. (2010). On the Role of Discourse Markers in Interactive Spoken Question Answering Systems. In *LREC*.
- Vasilescu, I., Hernandez, N., Vieru, B., & Lamel, L. (2018). Exploring Temporal Reduction in Dialectal Spanish: A Large-scale Study of Lenition of Voiced Stops and Coda-s. In *INTERSPEECH* (pp. 2728-2732).

OTELO


Identification/
selection of data
Oral database enriched with
multi-level metadata
Processing of homophones
and polysemic words
Multi-criteria phonetic
disambiguation of words
Extrapolate to big
data/heterogenous corpora
Testing results on effective
applications