Challenges of heterogeneous data for building Linguistic Theory

Anisia Popescu

anisia.popescu@universite-paris-saclay.fr

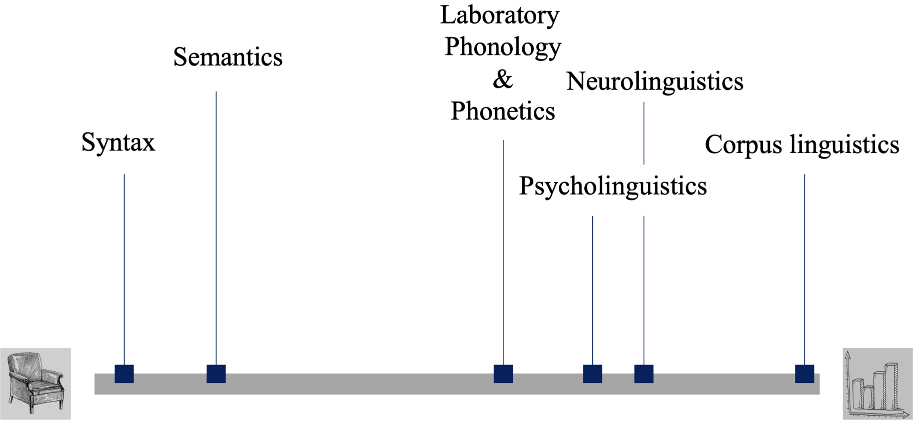
September 31 - October 3, 2024

"Every time I fire a linguist, the performance of the speech recognizer goes up"

F. Jelinek

1. Introduction
2. Heterogenous Data
3. Challenges for Linguistic Models
4. Discussion

Data in Linguistics



Syntax

Semantics

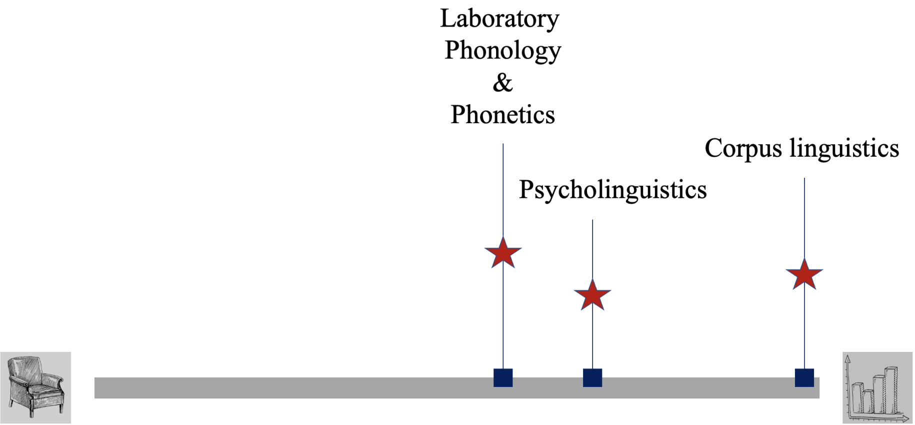
Laboratory Phonology & Phonetics

Psycholinguistics

Neurolinguistics

Corpus linguistics

Data in Linguistics

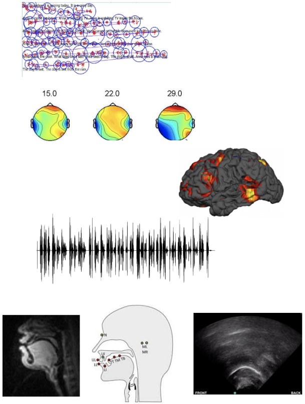


Laboratory Phonology & Phonetics

Psycholinguistics

Corpus linguistics

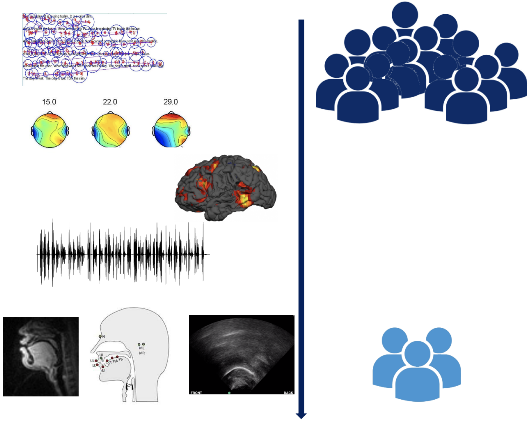
Data in Laboratory Linguistics



MULTIMODAL DATA

Image source: Rebernik etal., 2021, https://www.biopac.com/events/fmri-psych/, https://nilosarraf.com/

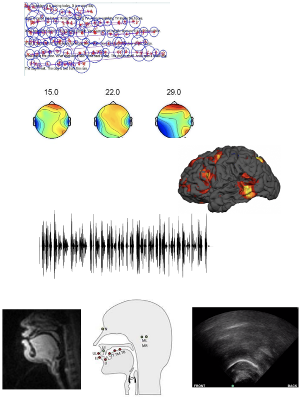
Data in Laboratory Linguistics



MULTIMODAL DATA

Image source: Rebernik etal., 2021, https://www.biopac.com/events/fmri-psych/, https://nilosarraf.com/

Data in Laboratory Linguistics



MULTIMODAL DATA

Highly curated corpora

Tailored to a specific RQ

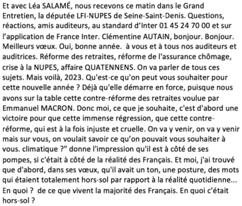
Limited participant pool

Low linguistics variation

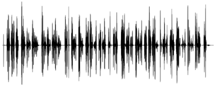
HETEROGENEITY?

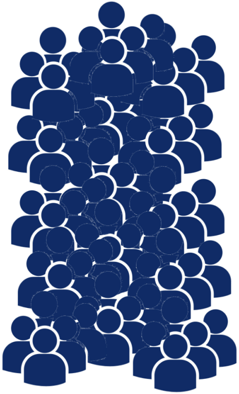
Image source: Rebernik etal., 2021, https://www.biopac.com/events/fmri-psych/, https://nilosarraf.com/

Data in Speech Technology



Et avec Léa SALAMÉ, nous recevons ce matin dans le Grand Entretien, la députée LFI-NUPES de Seine-Saint-Denis. Questions, réactions, amis auditeurs, au standard d'Inter 01 45 24 70 00 et sur l'application de France Inter. Clémentine AUTAIN, bonjour. Bonjour. Meilleurs vœux. Oui, bonne année. à vous et à tous nos auditeurs et auditrices. Réforme des retraites, réforme de l'assurance chômage, crise à la NUPES, affaire QUATENNENS. On va parler de tous ces sujets. Mais voilà, 2023. Qu'est-ce qu'on peut vous souhaiter pour cette nouvelle année ? Déjà qu'elle démarre en force, puisque nous avons sur la table cette contre-réforme des retraites voulue par Emmanuel MACRON. Donc moi, ce que je souhaite, c'est d'abord une victoire pour que cette immense régression, que cette contre-réforme, qui est à la fois injuste et cruelle. On va y venir, on va y venir mais sur vous, on voulait savoir ce qu'on pouvait vous souhaiter à vous. climatique ?" donne l'impression qu'il est à côté de ses pompes, si c'était à côté de la réalité des Français. Et moi, j'ai trouvé que d'abord, dans ses vœux, qu'il avait un ton, une posture, des mots qui étaient totalement hors-sol par rapport à la réalité quotidienne... En quoi ? de ce que vivent la majorité des Français. En quoi c'était hors-sol ?





Large sample size

Varied sources

Varied languages

Different speaking styles

MULTIMODAL DATA

HETEROGENEITY

1. Introduction
2. **Heterogenous Data**
3. Challenges for Linguistic Models
4. Discussion

Data in Speech Technology

HMM-GMM

control over acoustic parameterization

identifiable data

Deep Learning

no control

gate keeping

2010

Data in Speech Technology

Language Family	**Corpora**	**Duration**	**Speech style**
Romance	LDC, Quaero, Ester, Etape, NCCFR, Babel	1000	BN, BC, TC, C
Germanic	LDC, Quaero, Babel	1000	BN, C, TC
Slavic	LDC, Quaero, Babel	50-500 hours/language	BN, C, TC
Other	LDC, Quaero, Babel, Bulb	30+ hours/language	BN, C, TC

BN=broadcast prepared
BC=broadcast conversations
C=informal conversations
TC=telephonic conversations

Issues

- Data collected for ASR systems
    - Manual orthographic transcriptions
    - Automatically transcribed and manually corrected
- No linguistically formatted metadata
    - Lacking speaker diarization
        - no linguistic background information (dialect, native/non-native speaker, mono/bi/multi-lingual)
        - no diastratic factors (age, gender, social background, pathologies)
    - No linguistic annotation (POS, prosodic information etc.)

Sounds

Background music	Background noise	Mumbled speech	Laboratory speech	Denoised MRI speech
🔊	🔊	🔊	🔊	🔊

Case study

Phonetic typology and Language Evolution

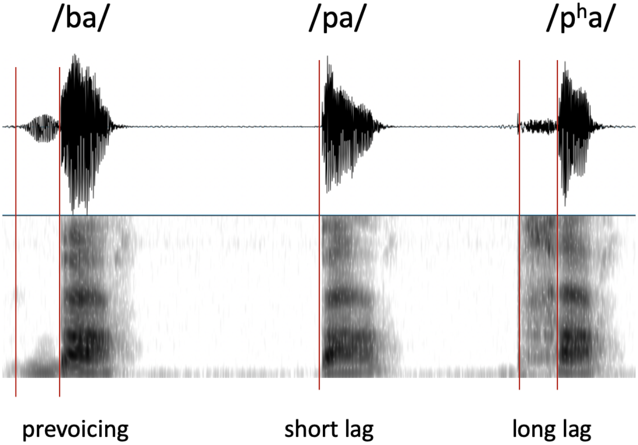
Research questions:

- How do linguistic behaviors emerge?
- How does language vary over time?

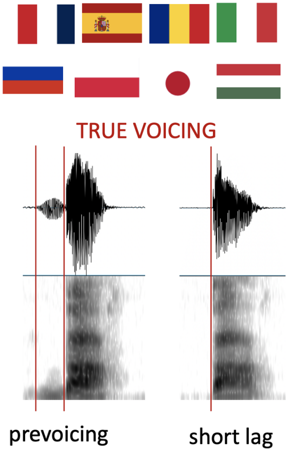
Case study:

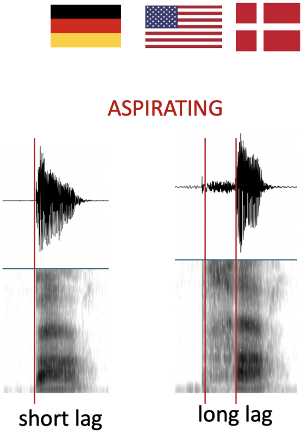
- Typological classification of Portuguese voicing system

Background



Background


TRUE VOICING
prevoicing
short lag


ASPIRATING
short lag
long lag

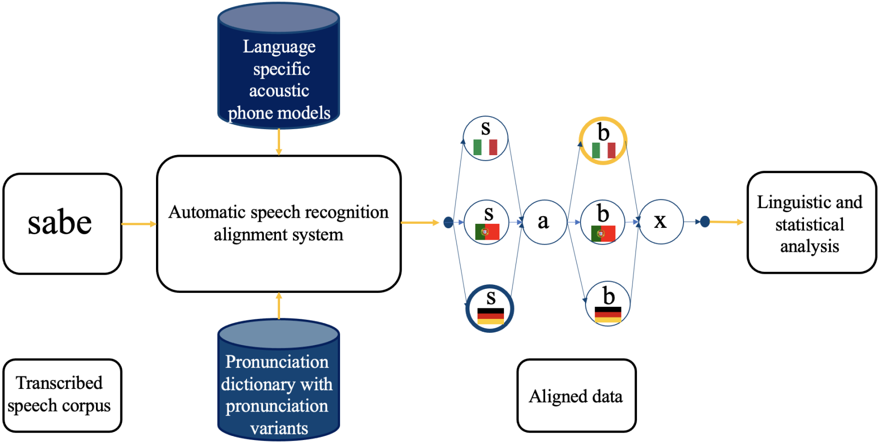
European Portuguese - the odd one out?

- European Portuguese - typically described as a voicing language
- Series of studies showing EP might be an aspirating language

  (Pape & Jesus, 2011, 2015, Shih & Möbius, 1999)
- One laboratory study suggesting EP might have a hybrid voicing system

  (Ramsammy & Strycharczuk, 2016)

Method



Language specific acoustic phone models

sabe

Automatic speech recognition alignment system

Linguistic and statistical analysis

Transcribed speech corpus

Pronunciation dictionary with pronunciation variants

Aligned data

Lamel etal. 2011

Acoustic models

Framework Lamel etal., 2011

- 3-state left-to-right continuous density HMM
- Gaussian mixtures with up to 32 Gaussians per state
- acoustic parametrization : cepstral (PLP) and pitch (F0) features
- word-, context-, speaker-independent monophone models

Training data

- EP models: 1.1 millions word tokens & 46k word types
- Italian acoustic models: 1.8 millions word tokens & 58.8k word types
- German acoustic models: 1.8 millions word tokens & 90k word types

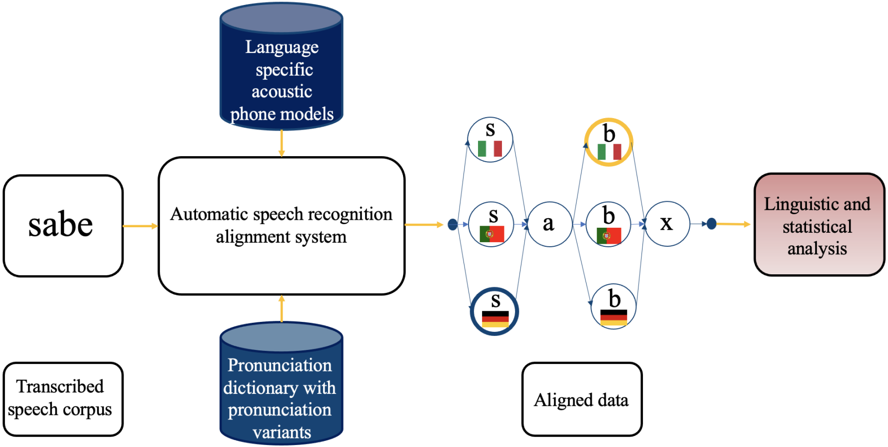
European Portuguese - the odd one out?

- Results support the proposal of a hybrid voicing system for EP obstruents (Ramsammy & Strycharczuk, 2016)
- Results could be an indication of sound change in European Portuguese: a shift form classic voicing systems known for Romance languages

BUT

- Results are more nuanced when adding phonological detail in the analysis
- The absence of speaker information limits the variance explained by statistical models.
- Manually coding speaker ID for a subset of the data results in higher R^2 (Cronenberg etal., 2024)

Method



sabe

Automatic speech recognition alignment system

Language specific acoustic phone models

Pronunciation dictionary with pronunciation variants

Transcribed speech corpus

Aligned data

Linguistic and statistical analysis

1. Introduction
2. Heterogenous Data
3. Challenges for Linguistic Models
4. **Discussion**

Way forward

Enhance the quality and consistency of linguistically formatted metadata

- Leverage publicly available sources
  - Knowledge Base for automatically annotating speech corpora (Wu etal. 2022), OTELO - Vasilescu & Suchanek, 2019
- Turn to related research domains that are more attuned to the specific metadata
  - Human and Social Sciences

Thank you for listening !

anisia.popescu@universite-paris-saclay.fr

Selected References

- Cronenberg, J., Lamel, L., Chitoran, I. (2024), Acoustic Assessment of the Diphthong vs. Hiatus Distinction in Five Romance Languages: A Big Data Study, Labphon19, Seoul, South Korea.
- Pape, D and Jesus, L. (2015). Stop and fricative devoicing in euorpean portuguese, italian and german. Language and Speech, 58(2):224–245.
- Popescu, A., Lamel, L., Vasilescu, A. (2024). Using Speech Technology to test Theories of Phonetic and Phonological Typology. In Proceedings of LREC-COLING 2024. Torino, Italy.
- Popescu, A., Hutin, M., Vasilescu, I., Lamel, L. and Adda-Decker. M. (2023). Stop devoicing and place of articulation: A cross-linguistic study using large-scale corpora. In Proceedings of the 20th International Congress of Phonetic Sciences, Prague, Czech Republic.
- Lamel, L. etal. (2011). Speech recognition for machine translation in quaero. In Proceedings of the 8th International Workshop on Spoken Language Translation: Evaluation Campaign, San Francisco, California.
- Ramsammy, M. and Strycharczuk, P. (2016). From phonetic enhancement to phonological underspecification: hybrid voicing contrast in european portuguese. Papers in Historical Phonology, 1.
- Wu, Y., Suchanek, F., Vasilescu, I., Lamel, L., Adda-Decker, M. (2022). Using a Knowledge Base to Automatically Annotate Speech Corpora and to Identify Sociolinguistic Variation. In LREC 2022.