La notion de réseau social en sociolinguistique computationnelle

Jean-Philippe Magué

Socionet

8 juin 2016

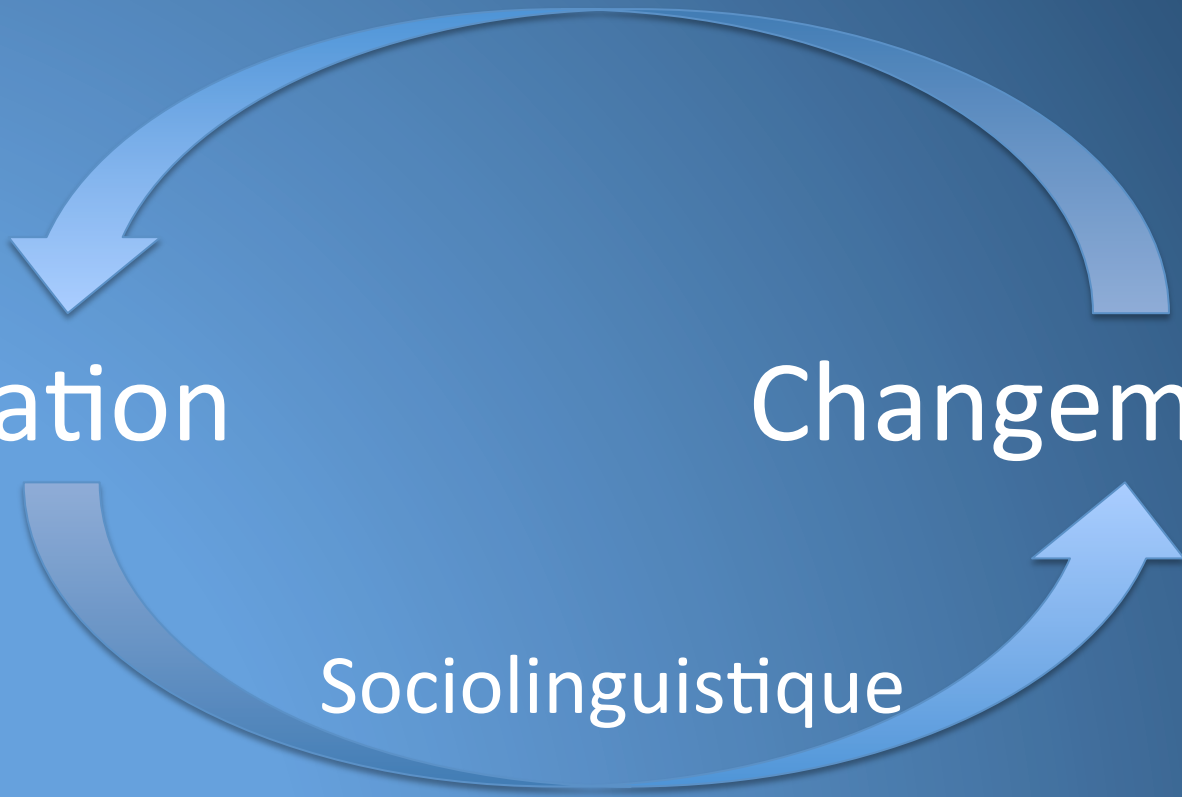
icar
UMR 5191 - CNRS

ENS DE LYON

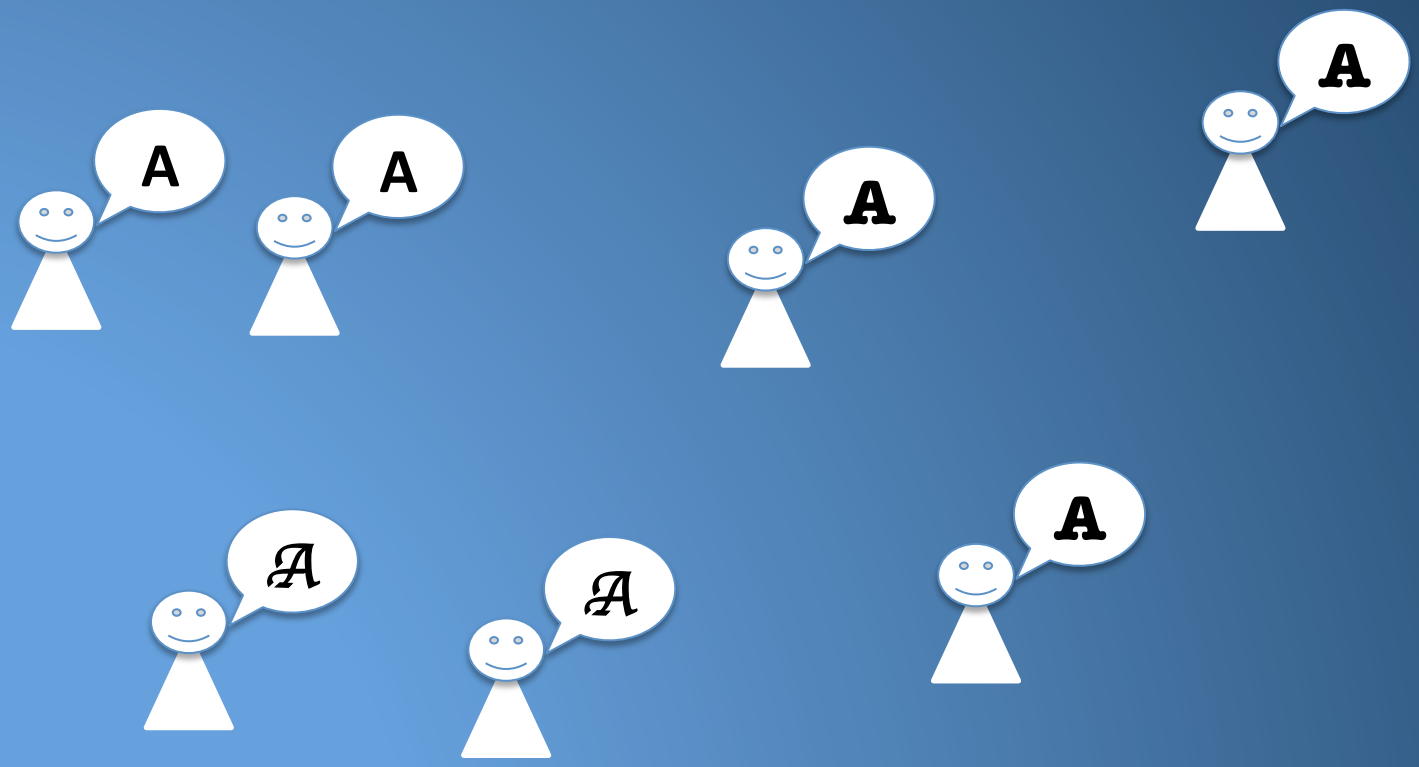
Les langues varient…

Atlas linguistique de France, Jules Guilleron, Edmond Edmont, 1902-1910

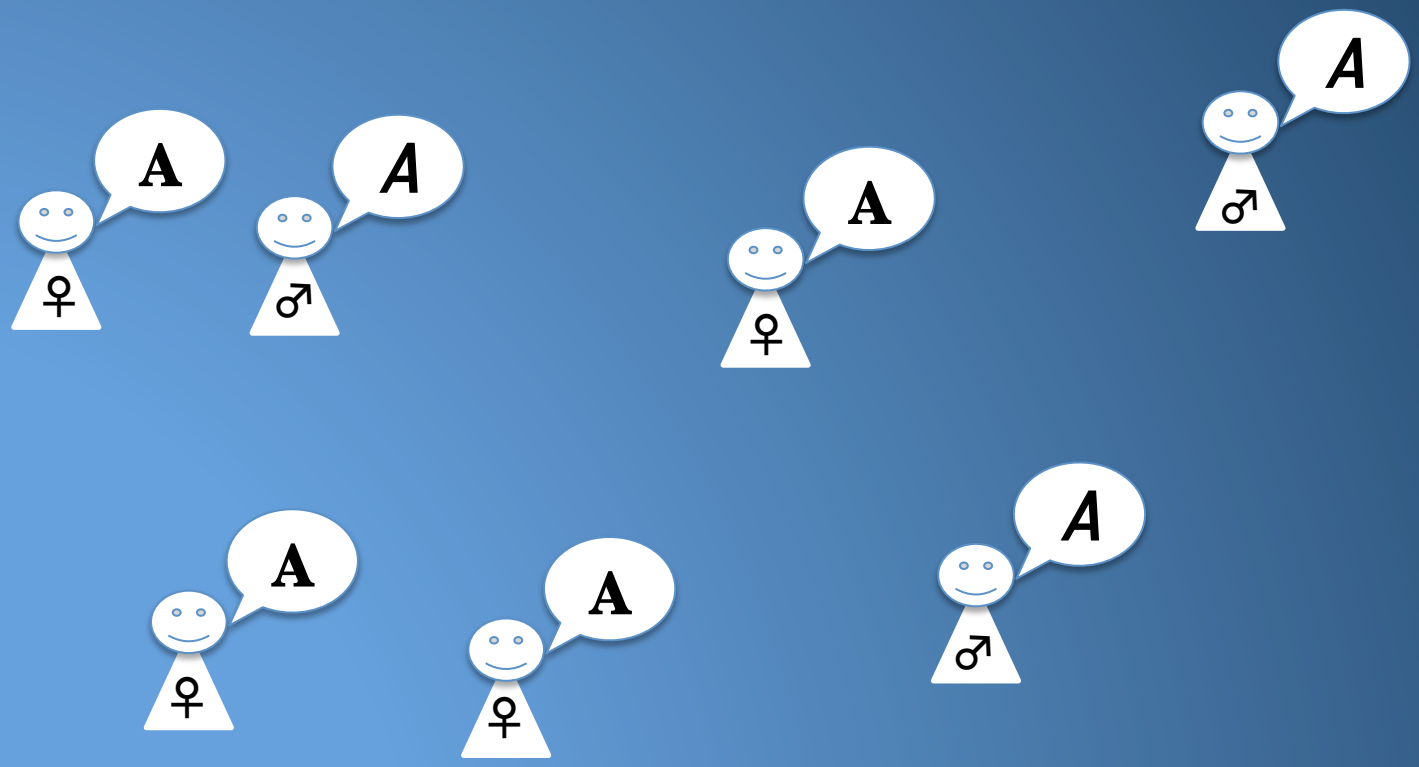
… et les langues changent



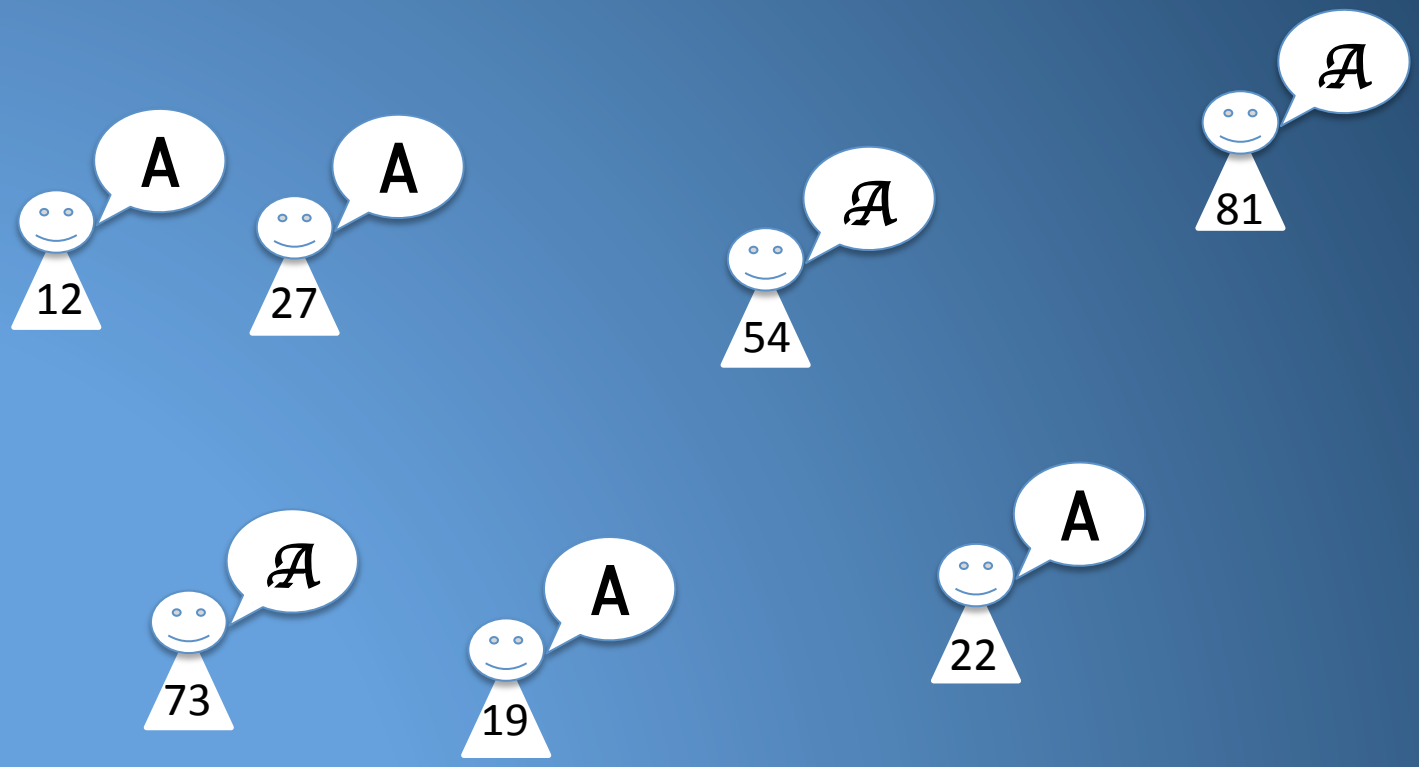
Sociolinguistique variationniste



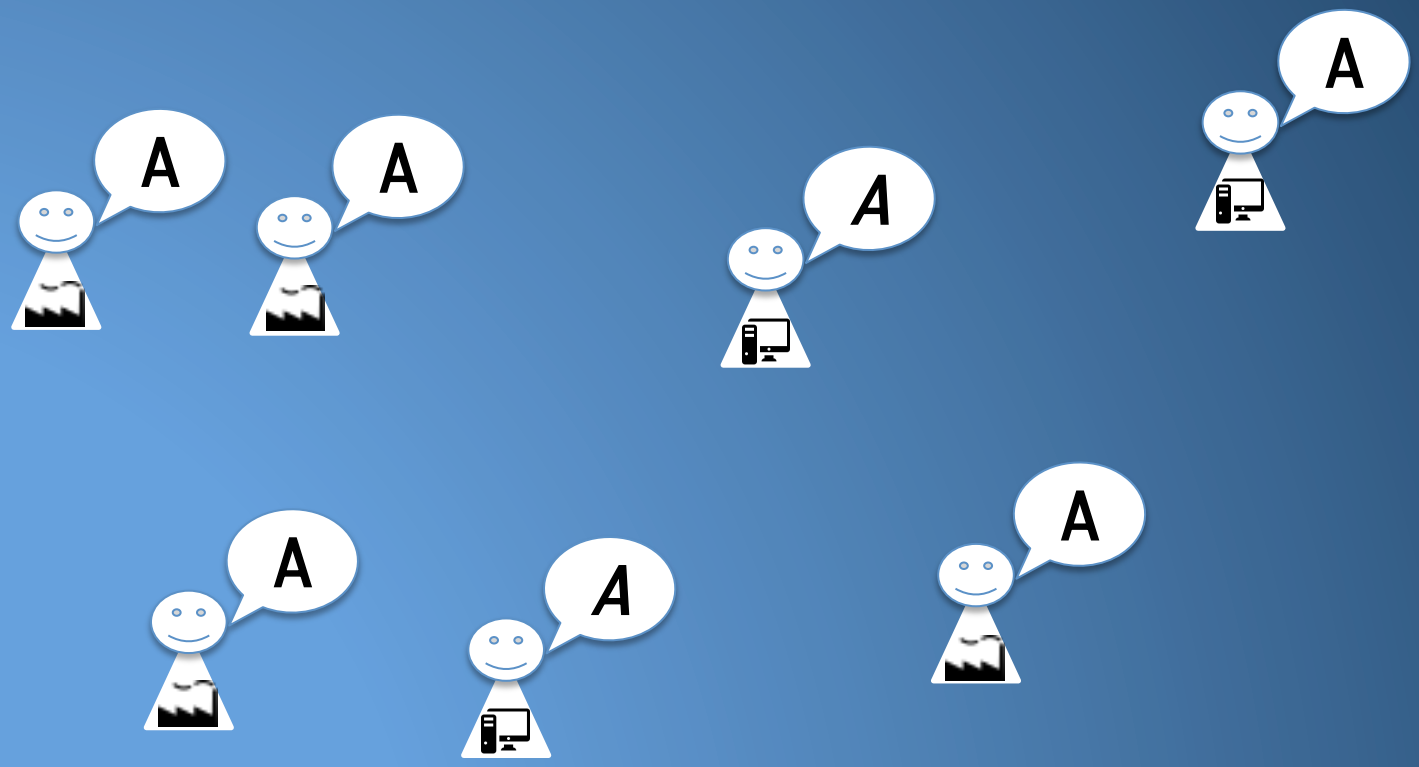
Sociolinguistique variationniste



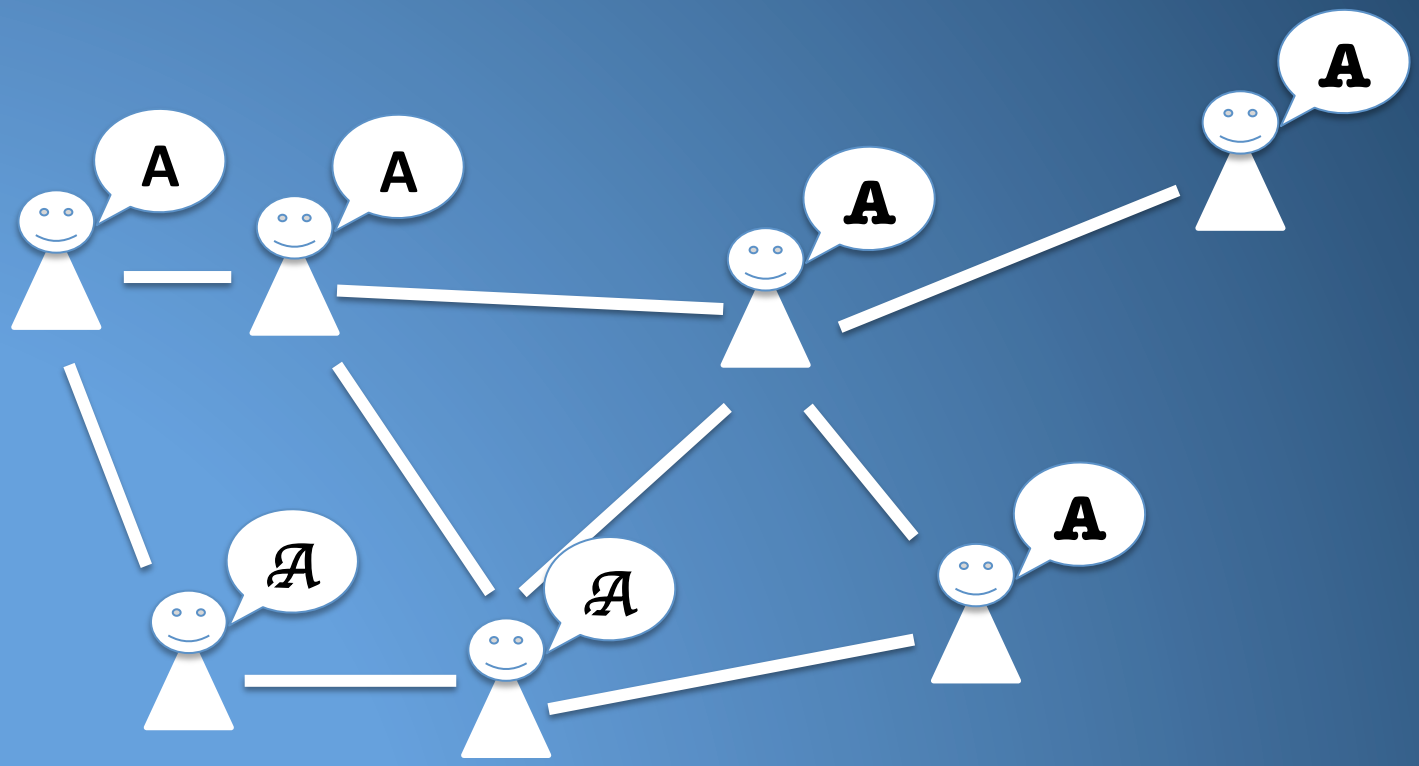
Sociolinguistique variationniste



Sociolinguistique variationniste



Sociolinguistique variationniste



Comment sait-on comment les gens parlent ?

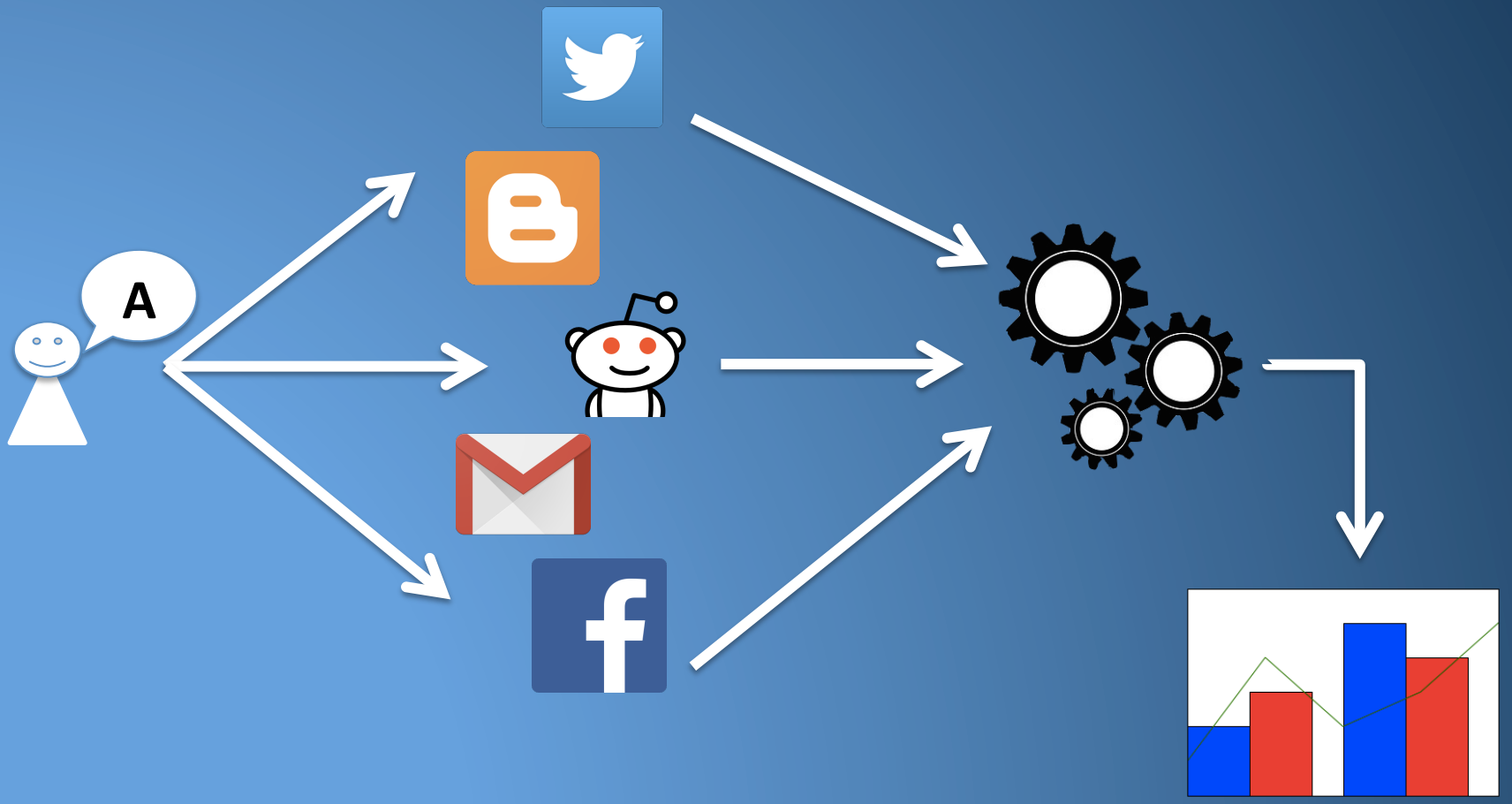
Cordereix Pascal, « Ferdinand Brunot et les Archives de la parole : le phonographe, la mort, la mémoire », Revue de la BNF 3/2014 (n° 48) , p. 5-11

Le paradoxe de l'observateur

« The aim of linguistic research in the community must be to find out **how people talk when they are not being systematically observed***; yet we can only obtain these data by systematic observation »*

Labov, 1972

Comment sait-on comment les gens parlent ?

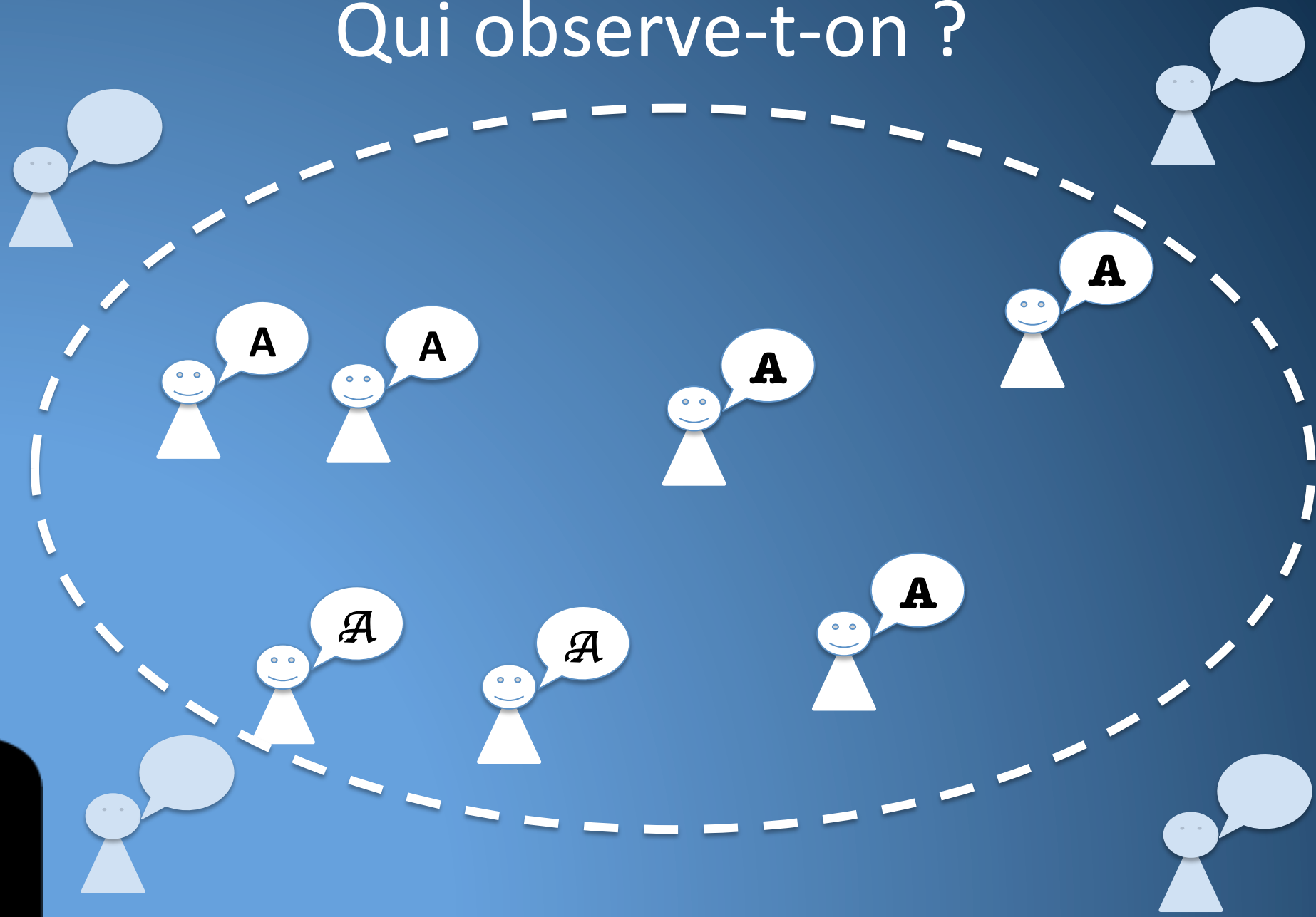


Qui observe-t-on ?

« The aim of linguistic research in the **community** *must be to find out how people talk when they are not being systematically observed; yet we can only obtain these data by systematic observation »*

Labov, 1972

Qui observe-t-on ?

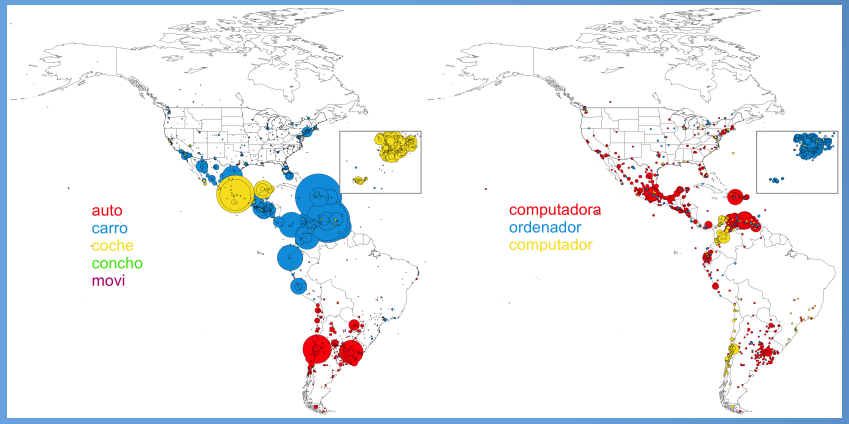


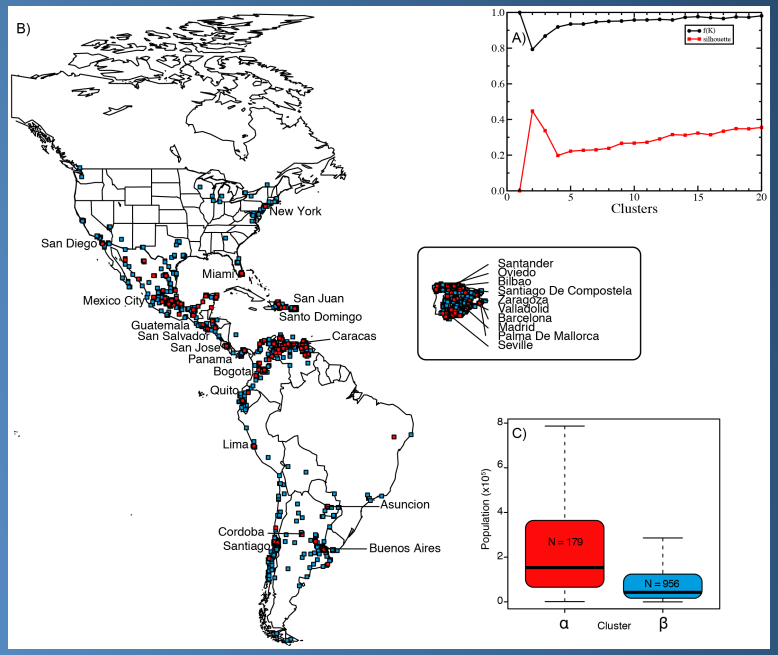
Une communauté, c'est une langue

« all the people who use a given language » Lyons, 1970

Crowdsourcing Dialect Characterization through Twitter
Gonçalves & Sánchez, 2014

50 millions de tweets en espagnol géolocalisés





Une communauté, c’est une langue

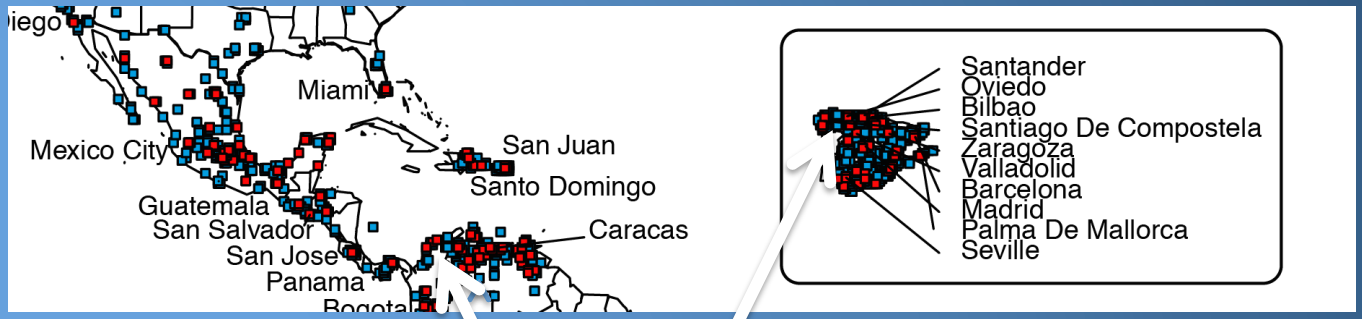
- Quelques problèmes
  - Qu’est-ce qu’une langue ?
    - Ce que reconnaît Chromium Compact Language Detector.
    - Sur le français, rappel de 67%

Une communauté, c'est une langue

- Quelques problèmes
  - Qu'est-ce qu'une langue ?
  - Circularité :
    - *communauté* pour caractériser la complexité/ hétérogénéité d'une langue.
    - Si la communauté est la langue…

Une communauté, c’est une langue

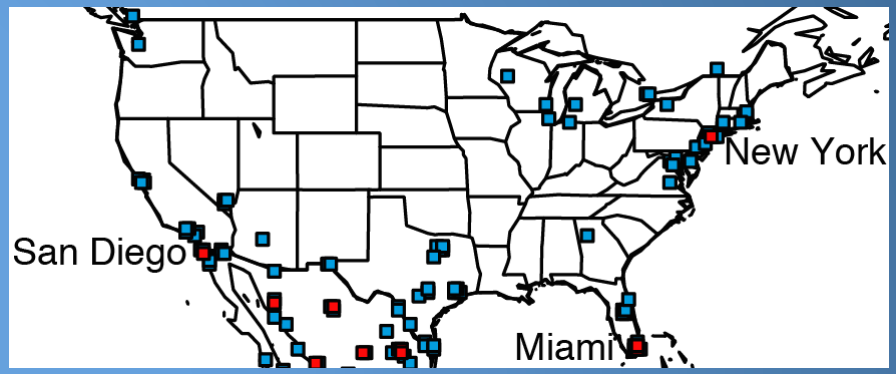
- Quelques problèmes
  - Qu’est-ce qu’une langue ?
  - Circularité
  - 1 même langue, deux communautés



?

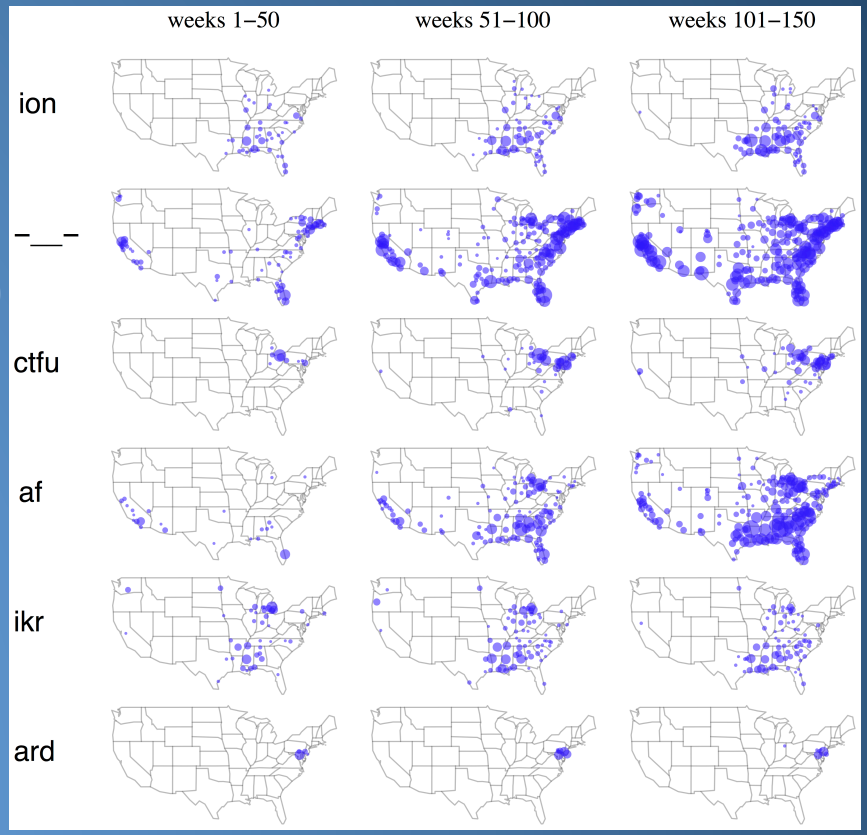
Une communauté, c'est une langue

- Quelques problèmes
  - Qu'est-ce qu'une langue ?
  - Circularité
  - 1 même langue, deux communautés
  - 2 langues, une même communauté



Une communauté se définit géographiquement

Diffusion of Lexical Change in Social Media
Eisenstein, O'Connor, Smith, Xing, 2014



107 millions de tweets géolocalisés
A l'intérieur des Etats-Unis
Assignés à une aire métropolitaine parmi 200

➔La diffusion d'une ville à l'autre est influencée par :

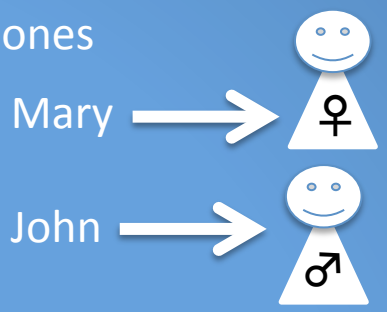
- La distance géographique
- La composition démographique
  - Afro-américains
  - Hispaniques

Une communauté se définit géographiquement

Gender identity and lexical variation in social media.
Bamman, Eisenstein, Schnoebelen, 2014.

10 millions de tweets
14000 utilisateurs:

- Aux US
- Anglophones



Word class	Previous literature	Our analysis
Pronouns	F	F
Emotion terms	F	F
Kinship terms	F	mixed
CMC words (*lol*, *omg*)	F	F
Conjunctions	F	ns
Clitics	F	ns
Articles	M	ns
Numbers	M	M
Quantifiers	M	ns
Technology words	M	M
Prepositions	mixed	ns
Swear words	mixed	M
Assent	mixed	F
Negation	mixed	mixed
Emoticons	mixed	F
Hesitation	mixed	F

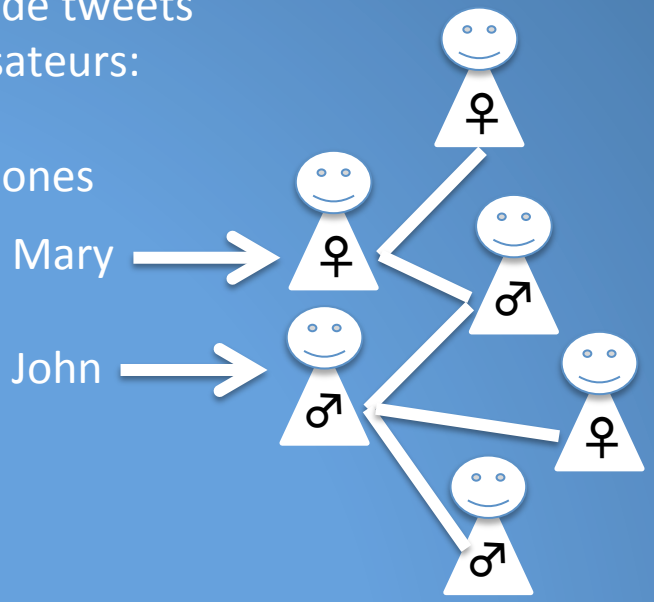
➔Des spécificités linguistiques propres à chaque genre

Une communauté se définit géographiquement

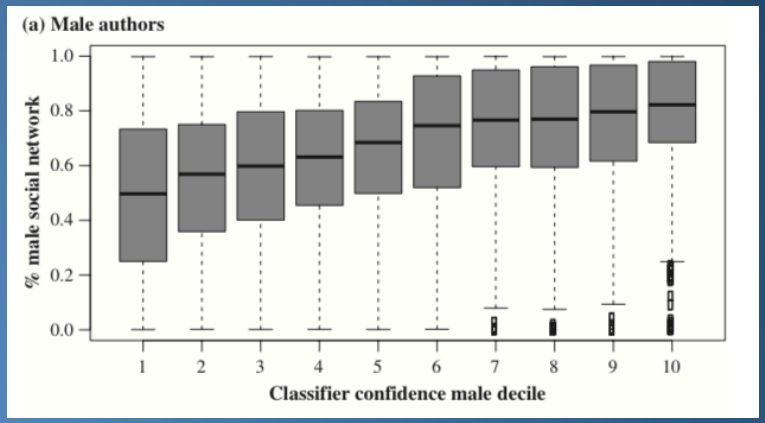
Gender identity and lexical variation in social media.
Bamman, Eisenstein, Schnoebelen, 2014.

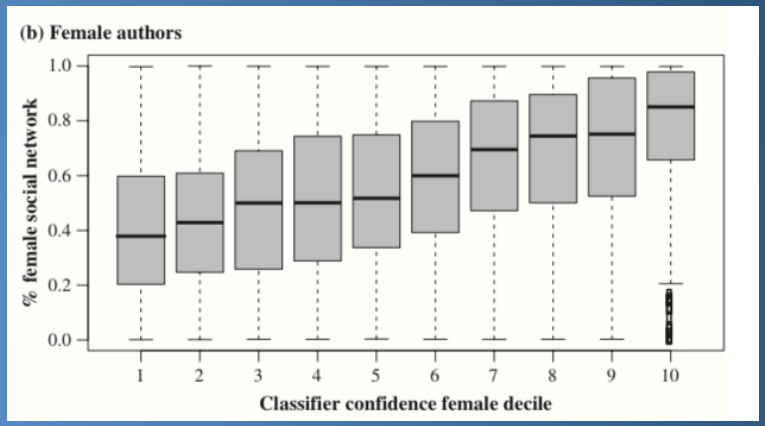
10 millions de tweets
14000 utilisateurs:

- Aux US
- Anglophones



➔L'homophilie entraine la proximité linguistique





La communauté, c'est le partage d'un système de normes

The speech community is not defined by any marked agreement in the use of language elements, so much as by participation in a set of shared norms Labov, 1972

La communauté, c'est le partage d'un système de normes

Language use as a reflection of socialization in online communities
Nguyen & Rosé, 2011

Forum sur le cancer du sein
3000 utilisateurs
8 ans

Figure 1: KL divergence between weeks in first year and overall second year.

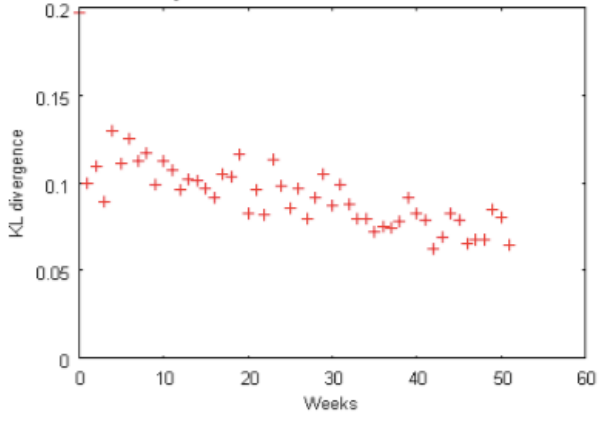
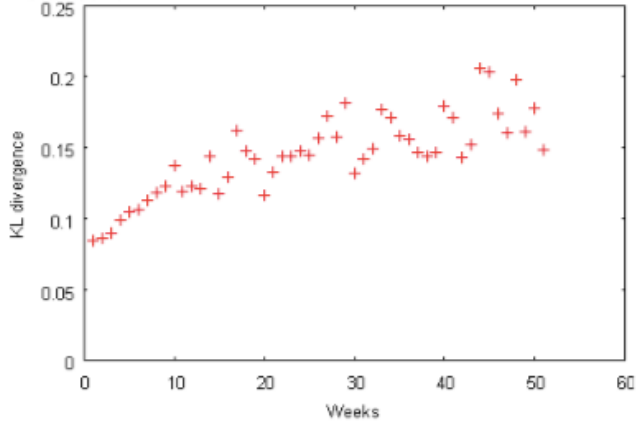


Figure 2: KL divergence between weeks in first year and first week.



➔Convergence vers la norme de la communauté

Communautés de pratique

*« an aggregate of **people who come together around mutual engagements** in some common endeavor. Ways of doing things, ways of talking, beliefs, values, power relations – in short, practices – **emerge** in the course of their joint activity around that endeavor.»*

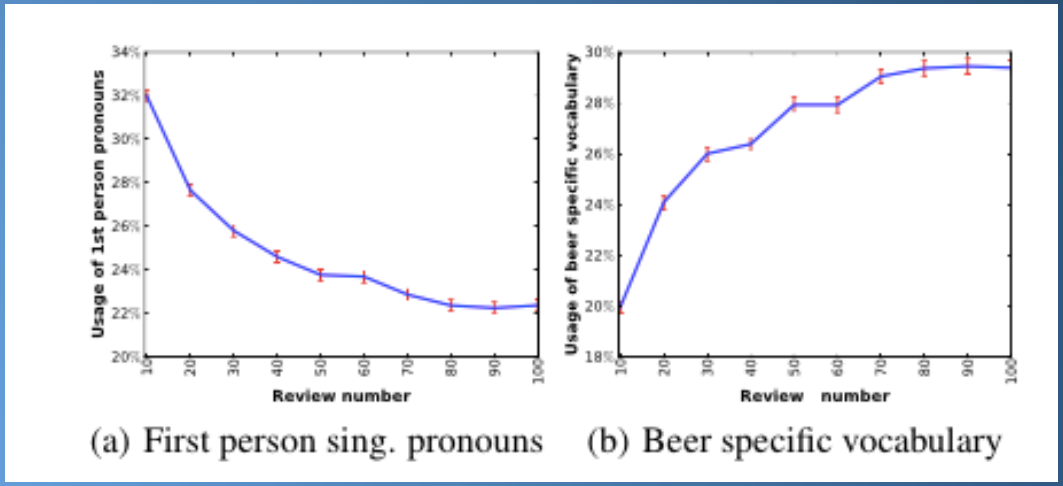
Eckert & McConnell-Ginet, 1998

Communautés de pratique

No Country for Old Members: User Lifecycle and Linguistic Change in Online Communities

Danescu-Niculescu-Mizil et al, 2013

2 sites d'avis sur les bières
10 ans
10000 utilisateurs



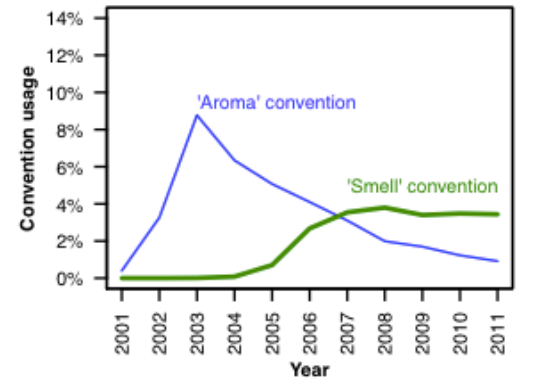
(a) First person sing. pronouns (b) Beer specific vocabulary

➔Convergence vers la norme de la communauté

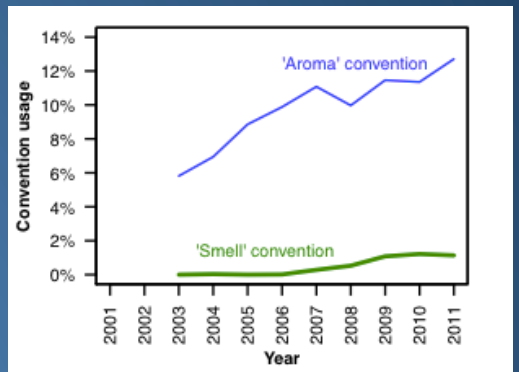
Communautés de pratique

No Country for Old Members: User Lifecycle and Linguistic Change in Online Communities

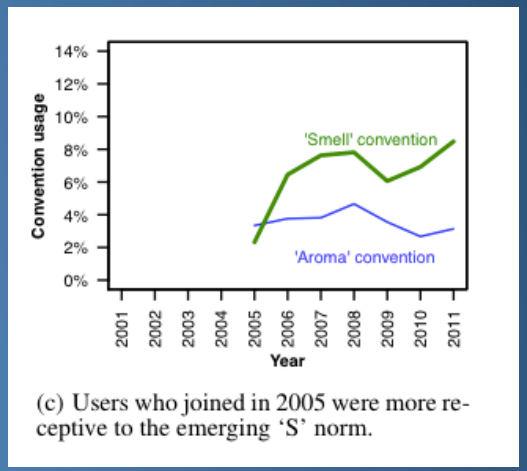
Danescu-Niculescu-Mizil et al, 2013



(a) 'Aroma' was the dominant convention by 2003, but it was supplanted by 'S' (for 'Smell') around 2007.



(b) Users who joined in 2003 hung on to the 'Aroma' convention of their youth.



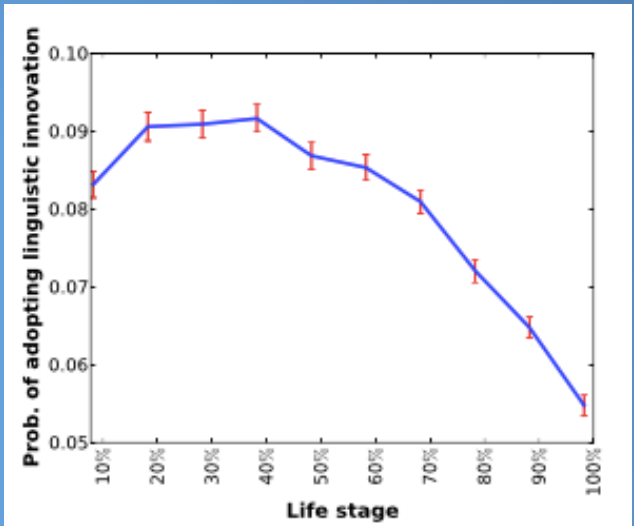
(c) Users who joined in 2005 were more receptive to the emerging 'S' norm.

➔Evolution de la norme de communauté

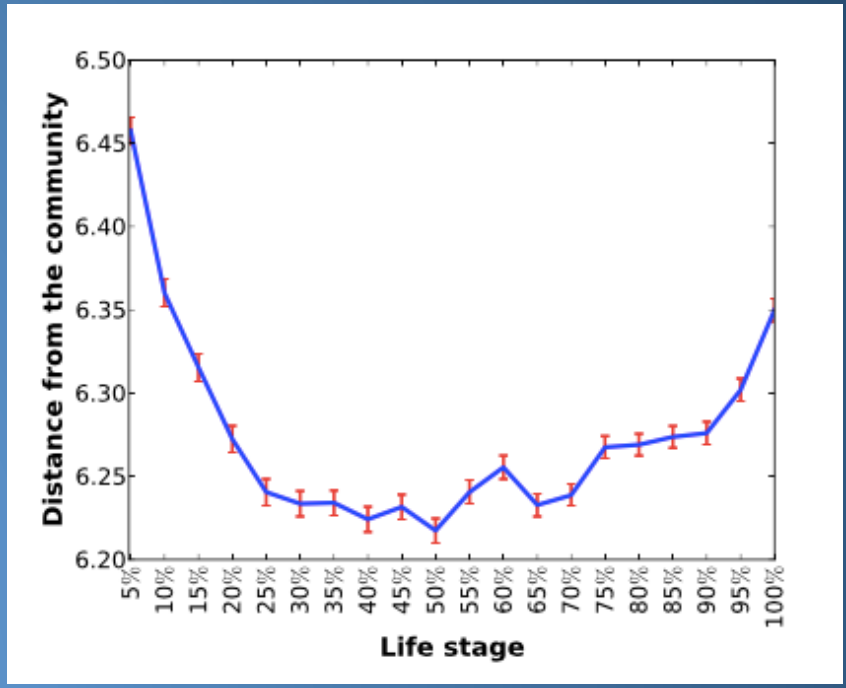
Communautés de pratique

No Country for Old Members: User Lifecycle and Linguistic Change in Online Communities

Danescu-Niculescu-Mizil et al, 2013



(c) Adoption of lexical innovations



➔On peut devenir vieux très rapidement

Une communauté, c'est des gens qui se parlent entre eux

a social group which may be either monolingual or multilingual, **held together by frequency of social interaction patterns** and **set off from the surrounding areas by weaknesses in the lines of communication**.

Gumperz 1971

➔Modularité en network science

Twitter : Oral ou écrit ?

Les distributions des parties du discours diffèrent entre l'oral et l'écrit

L'écrit penche plus vers les noms

la fatigue augmente avec la durée de la course

L'oral penche plus vers les verbes

plus on court, plus on est fatigué

"*Written language presents phenomena as if they were products. Spoken language presents phenomena as if they were processes*."
(Halliday, 1994)

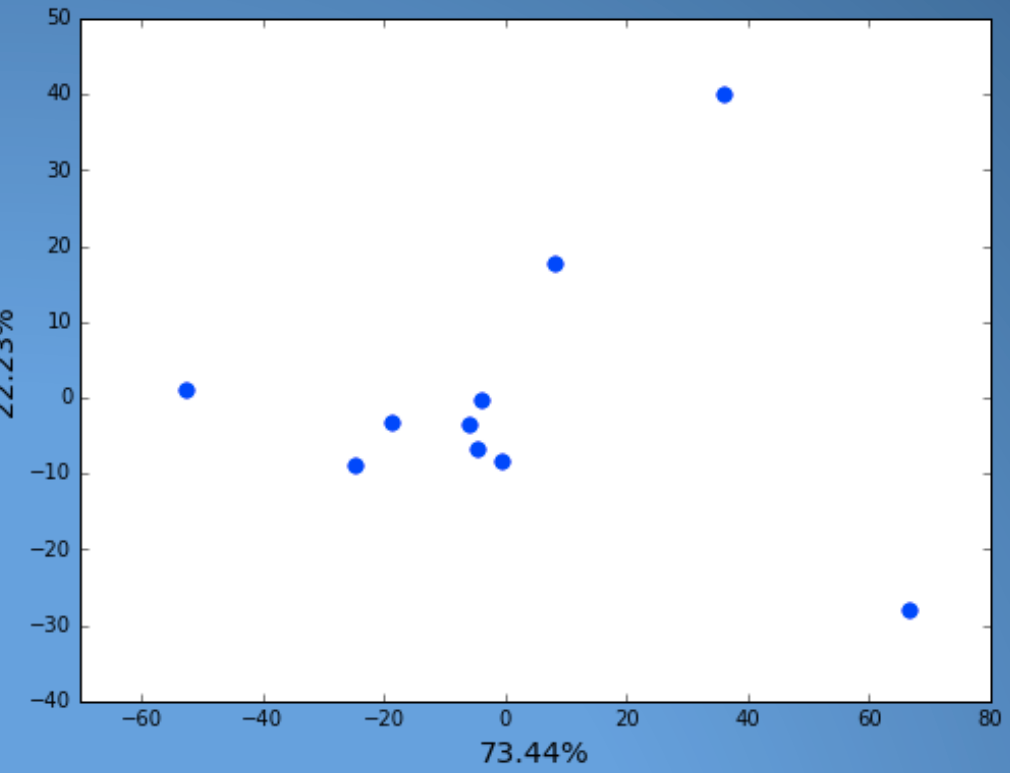
Twitter : Oral ou écrit ?

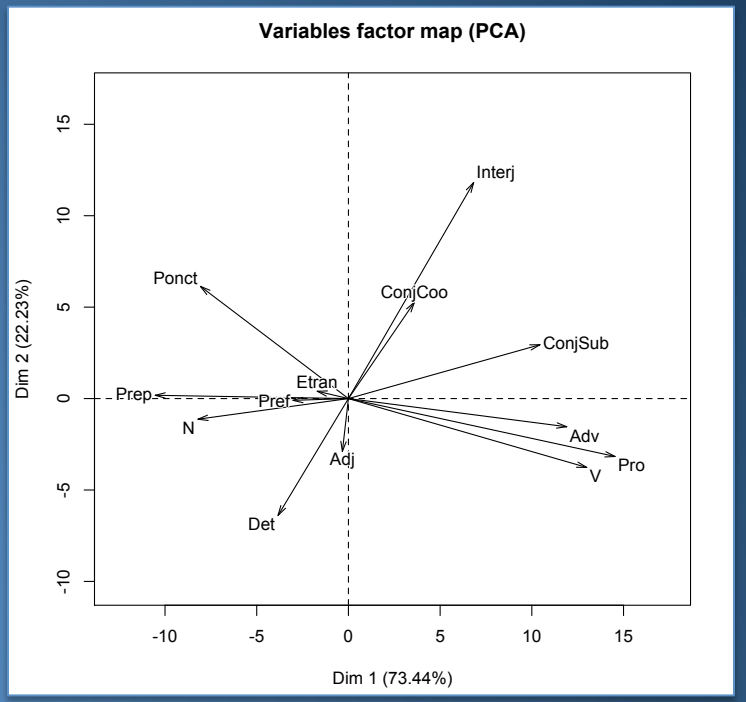
70 millions de tweets
- 10% des tweets en français entre juin 2014 et juin 2015
- Annotés en parties du discours (13 étiquettes)

Réseau de 1.7 million d'utilisateurs
- Qui suit qui
- 10 plus grosses communautés

Matrice M 10x13
- M_{ij} : sur- ou sous-réprésentation de la PDD *j* dans la communauté *i*
- Analyse en composantes principales

Twitter : Oral ou écrit ?





➔Les communautés se repartissent le long d'un continuum écrit / oral

Conclusion

- Les populations sont hétérogènes
- L'hétérogénéité
  - Multidimensionnelle
  - Structurée : il y a des formes d'homogénéité locale
  - Dynamique
  - Intriquée : il y a des corrélations/causations entre les différentes modalités

Conclusion

- Sociolinguistique :
  - lien entre l'hétérogénéité structurée des langues et d'autres d'hétérogénéité structurée sociale
  - S’appuie la notion de communauté
- Notion polymorphe
  - Différents buts, différentes approches

Conclusion

Most groups of any permanence, be they small bands bounded by face-to-face contact, modern nations divisible into smaller subregions, or even occupational associations or neighborhood gangs, may be treated as speech communities, provided they show linguistic peculiarities that warrant special study

Gumperz, 1971

Conclusions

- Approches computationnelles :
  - Renouvellement radical des données
  - Renouvellement radical des analyses
  - Assez conservatrices sur la définition de communauté
- La sciences des réseaux pourvoyeuse de nouveaux regards sur les relations individu/ population ?

Merci de votre attention

ANR-15-CE38-0011-01