Towards automatic geolocalisation of speakers of European French

Scherrer, Yves; Goldman, Jean-Philippe

Towards automatic geolocalisation of speakers of European French

Yves Scherrer & Jean-Philippe Goldman
University of Geneva

Automatic speaker geolocalisation

Data

Simulation and methods :

- Clustering and shibboleth detection
- Recursive feature elimination

Crowdsourced results

Automatic speaker geolocalisation

Ask a speaker *n* questions and predict his/her most likely area of origin (one out of *m* areas) with *p*% accuracy.

Automatic speaker geolocalisation

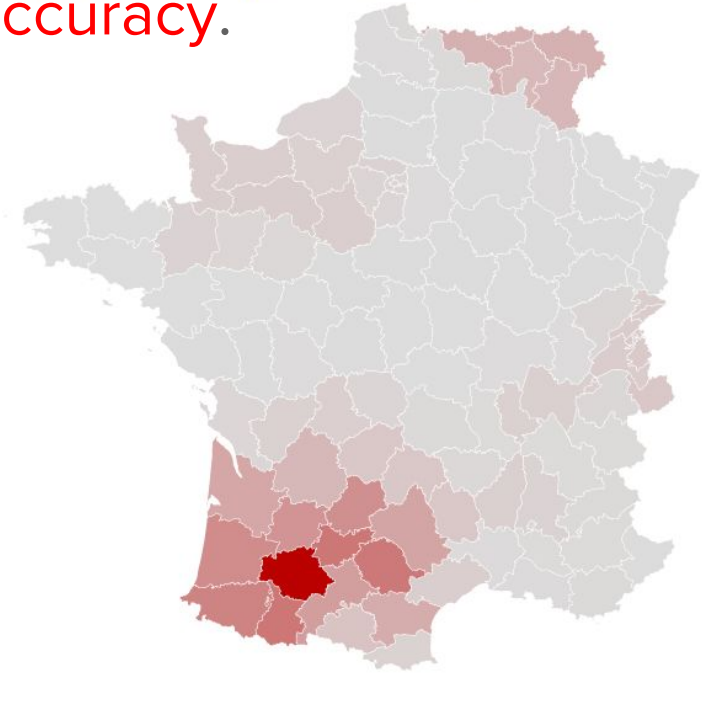
Ask a speaker *n* questions and predict his/her most likely area of origin (one out of *m* areas) with *p*% accuracy.

Comment appelez-vous cette pâtisserie ?

- [ ] Pain au raisin
- [ ] Escargot
- [ ] Cagouille
- [ ] Schnäcke
- [ ] Alsacienne
- [ ] Pain russe

Automatic speaker geolocalisation

Ask a speaker *n* questions and predict his/her most likely area of origin (one out of *m* areas) with *p*% accuracy.



Automatic speaker geolocalisation

Ask a speaker *n* questions and predict his/her most likely area of origin (one out of *m* areas) with *p*% accuracy.

Goals:

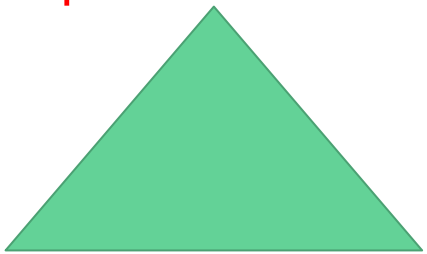
- Provide a playful incentive to attract participants for further inquiries
- Collect more data
- Observation ➔ Prediction
- Explore scientific analysis methods of the already collected data

⇒ select questions and areas to maximize accuracy

Automatic speaker geolocalisation

Ask a speaker *n* questions and predict his/her most likely area of origin (one out of *m* areas) with *p*% accuracy.

Expected accuracy of predictions ↗↗



Number and type of questions asked ↘↘

Number and type of predicted areas □□

Automatic speaker geolocalisation

Previous work:

- Create a geolocalisation model using data **from atlases**
- Select *n* questions on the basis of a **dialectologist's knowledge**
- Use the same *m* areas as in the **original data**
- Assess accuracy post-hoc (compare model predictions with participants' real origins)

(Leemann since 2013)

(parlometre.ch - TSR - 2015)

Automatic speaker geolocalisation

Previous work:

- Create a geolocalisation model using data **from atlases**
- Select *n* questions on the basis of a **dialectologist's knowledge**
- Use the same *m* areas as in the **original data**

- Assess accuracy post-hoc (compare model predictions with participants' real origins)

Our approach:

- … **from online inquiries**

- Select **optimal** *n* questions **by statistics**
- Select optimal *m* areas **by statistics**
- Estimate accuracy (given n and m) using the same data as for model creation and
- Assess accuracy post-hoc, **compare with estimates**

Data

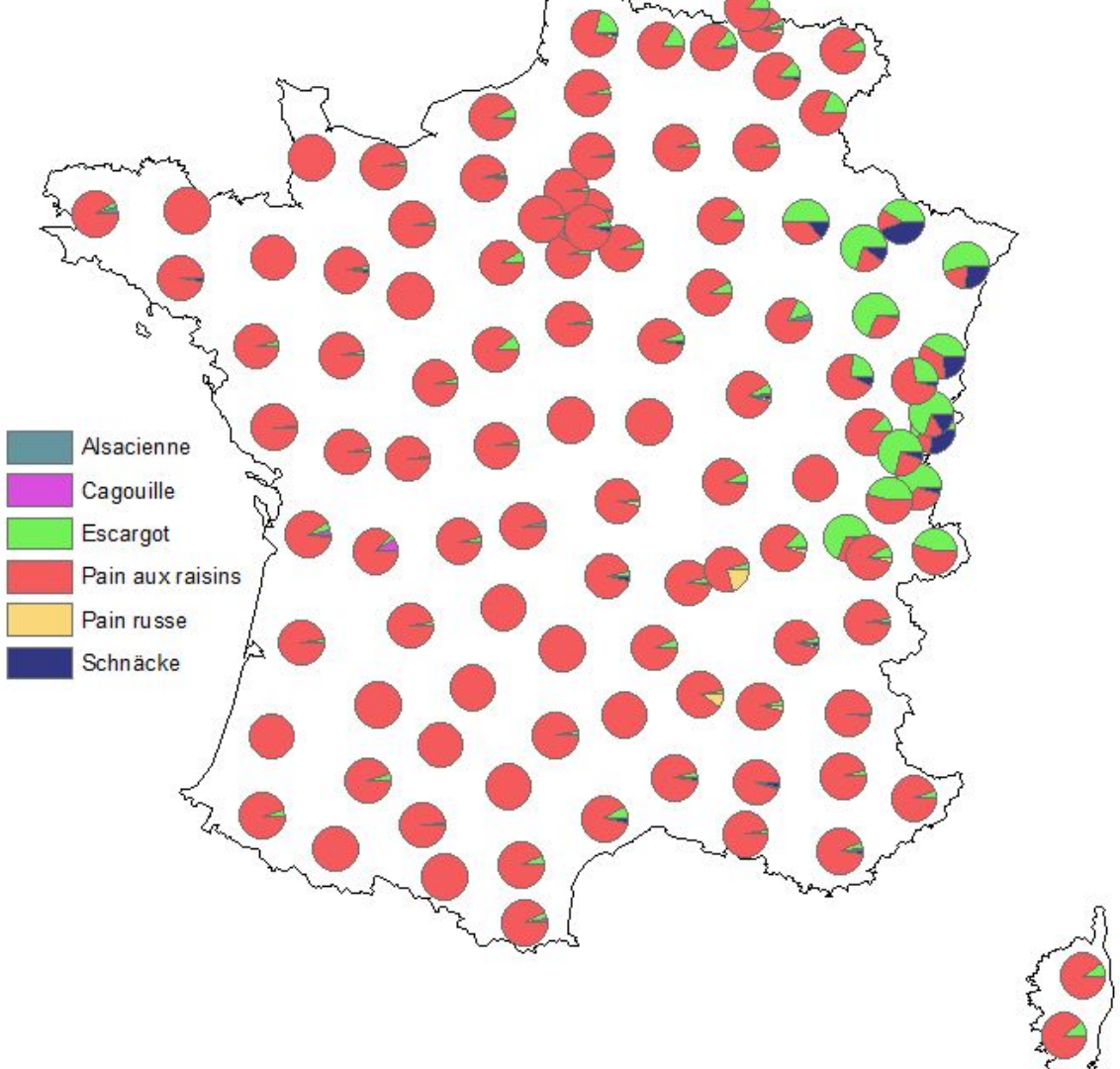
Project *Français de nos régions* (Avanzi, Glikman et al., 2015)

➜ online surveys to inquire about regionalisms in European French (France, Belgium, Switzerland).

Survey 1	**Survey 2**
May 2015 - May 2016	September 2015 - May 2016
40 questions	90 questions
12 000 participants	8 000 participants

Comment appelez-vous cette pâtisserie ?

- [ ] Pain au raisin
- [ ] Escargot
- [ ] Cagouille
- [ ] Schnäcke
- [ ] Alsacienne
- [ ] Pain russe
- [ ] Autre (précisez) :



Simulation

Simulation framework: {questions} + {areas} ➔ prediction accuracy

Idea: Leave-one-out method using two views of the same dataset

- Train model on aggregated data of all except one participant
- Predict origin of left-out participant, compare to ground truth

We do not leave out the test participant from the aggregated data:

- Much faster, as we don’t have to train a new model for each participant
- Since training data are aggregated and there are always > 1 participants per area, there is never an exact correspondence between training and test data
- Preliminary tests show good correlation with true leave-one-out method

Simulation

Simulation framework: {questions} + {areas} ➔ prediction accuracy

Two preprocessing steps:

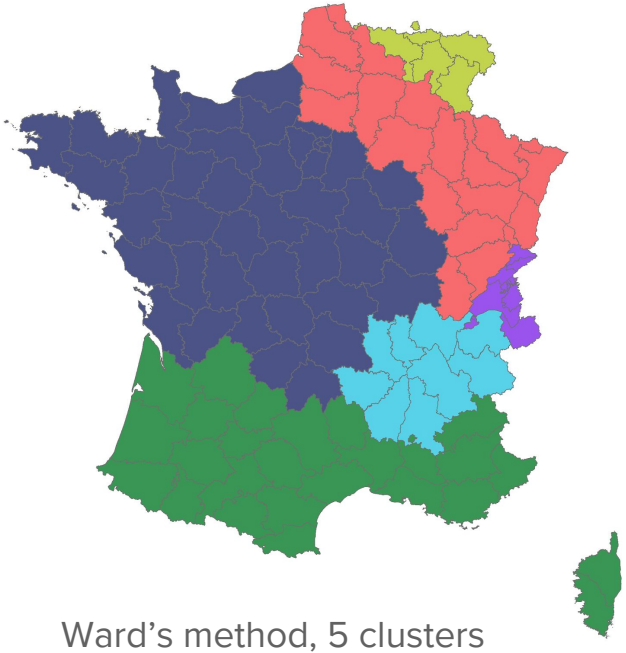
1. Settle on initial set of areas: FR départements, BE provinces, CH cantons (110)
2. Match participants from Survey 1 with participants from Survey 2 (same origin)

Two approaches to find {questions} and {areas}:

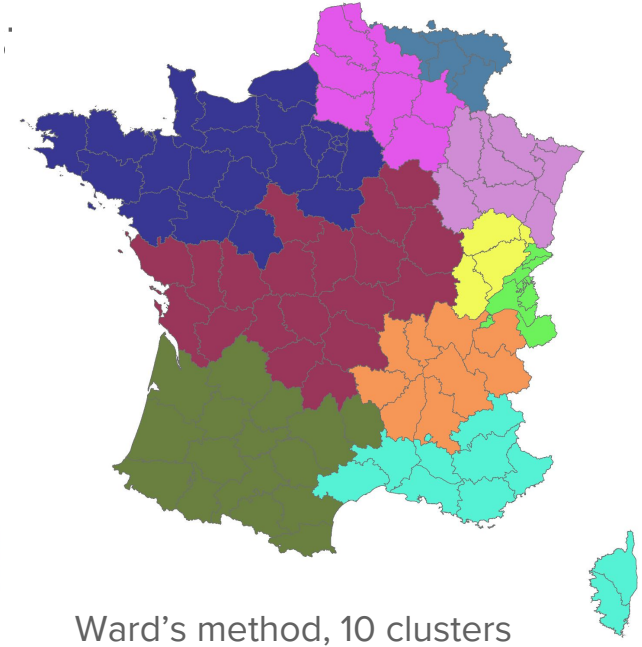
1. Clustering and shibboleth detection
2. Recursive feature elimination

Clustering and shibboleth detection

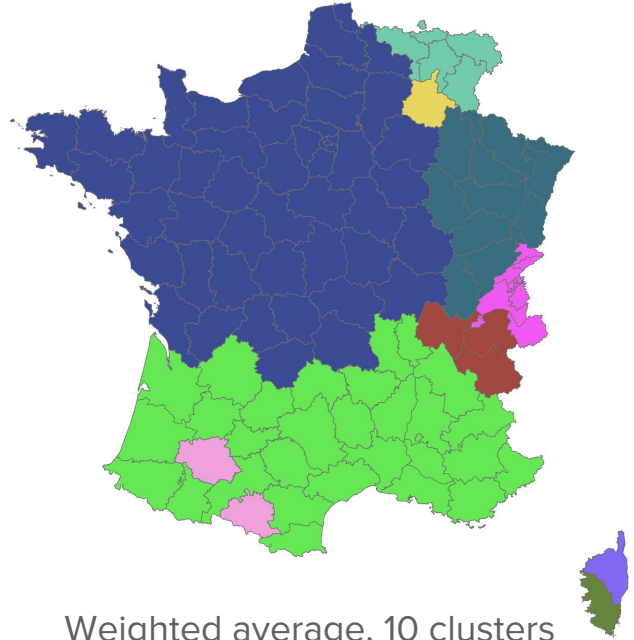
1. Determine the most relevant areal partition using



Ward's method, 5 clusters



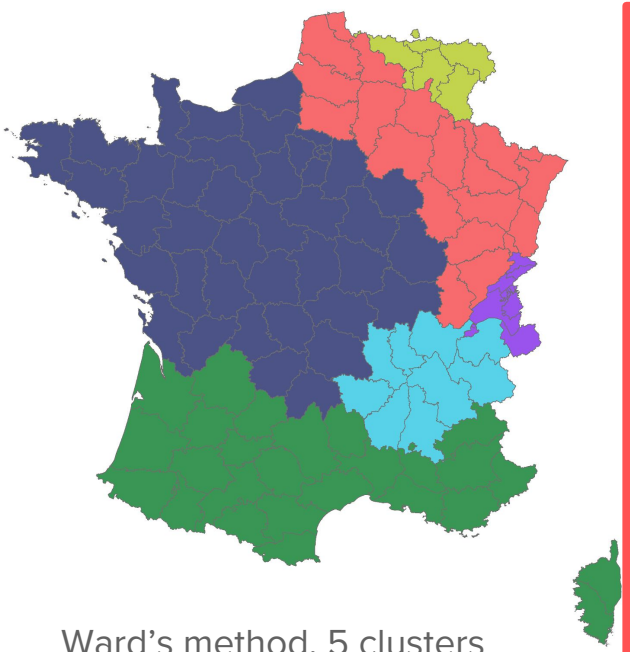
Ward's method, 10 clusters



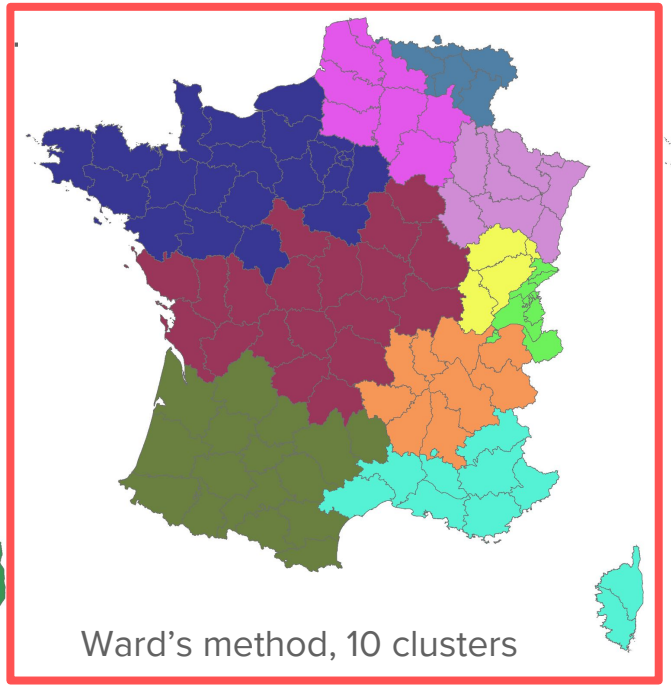
Weighted average, 10 clusters

Clustering and shibboleth detection

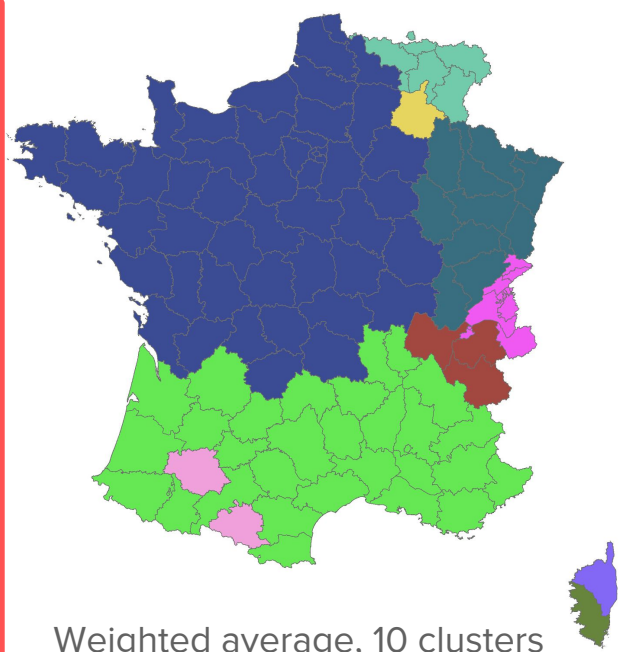
1. Determine the most relevant areal partition using



Ward's method, 5 clusters



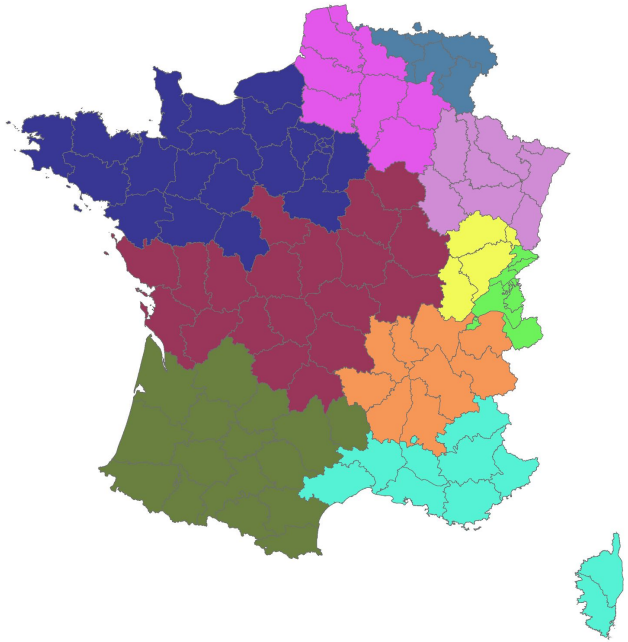
Ward's method, 10 clusters



Weighted average, 10 clusters

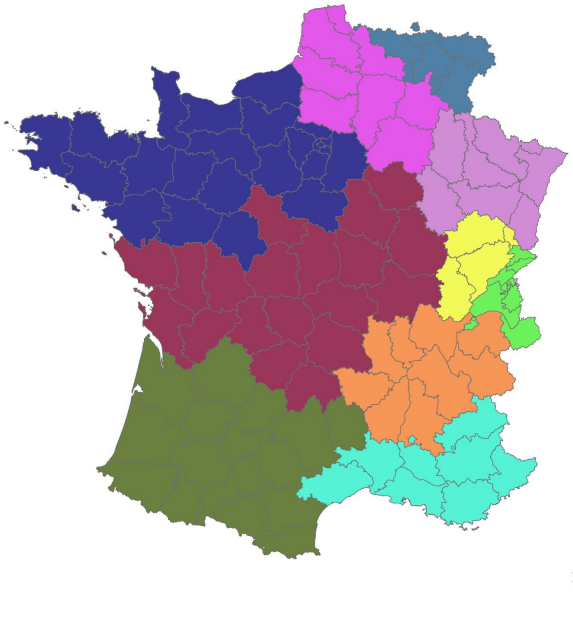
Clustering and shibboleth detection

2. Use the **shibboleth** detection algorithm (Prokic, Çöltekin & Nerbonne 2012) to find the most characteristic questions for each area (e.g. 5 shibboleths/cluster)



Clustering and shibboleth detection

2. Use the **shibboleth** detection algorithm (Prokic, Çöltekin & Nerbonne 2012) to find the most characteristic questions for each area (e.g. 5 shibboleths/cluster)



Morve
Quatre-vingt-dix
Soixante-dix
Ving(t)
Sèche-cheveux

Soixante-dix
Sèche-cheveux
Quatre-vingt-dix
Morve
Groseilles

Soixante-dix
Quatre-vingt-dix
Sèche-cheveux
Chocolatine
Groseilles

Sèche-cheveux
Groseillles
Clignotant
Quatre-vingt-dix
Soixante-dix

Groseilles
Sèche-cheveux
Clignotant
Sécher
Nombril

Péguer
Challer
Soixante-dix
Sèche-cheveux
Quatre-vingt-dix

Essuie-tout
Septante
Nonante
Quelle heure il-est?
Morve

Débarouler
Sèche-cheveux
Ving(t)
Groseilles
Clignotant

Soixante-dix
Quatre-vingt-dix
Groseillles
Flaques
Clignotant

Encoubler/Achouper
Septante
Nonante
Ca joue
Souper

Clustering and shibboleth detection

Simulation results:

- 10 clusters, all 130 questions ➜ 65.1% correct
  - The results are very sensitive to the cluster borders:
    -24% between 4 and 5 clusters; -21% between 10 and 11 clusters
  - It is difficult to determine a “good” number of clusters and an optimal cluster algorithm
- 10 clusters, 14 manually defined questions ➜ 67.0% correct
  - Few carefully selected questions are better than all questions
- 10 clusters, 20 questions determined by shibboleth detection ➜ 61.8% correct
  - Unintuitive choice of questions (standard variants for most areas)
  - Clusters are defined on all data, not on single determining questions

Recursive feature elimination

1. The linguistic variables may have several variants with different distributions. Treat each variant separately.
2. Some variants are hardly ever used or show no geographic variation at all. Discard them first.
3. Train a classifier with the remaining variants, remove the one variant that contributes least to the classification, repeat.
4. Use the 110 atomic areas and distance between centroids throughout the process. At the end, dynamically extend the areas to their immediate and second-order neighbors.

Recursive feature elimination

1. The linguistic variables may have several variants with different distributions. Treat each variant separately.

Binarize data: 130 n-ary variables ➜ 639 binary variables

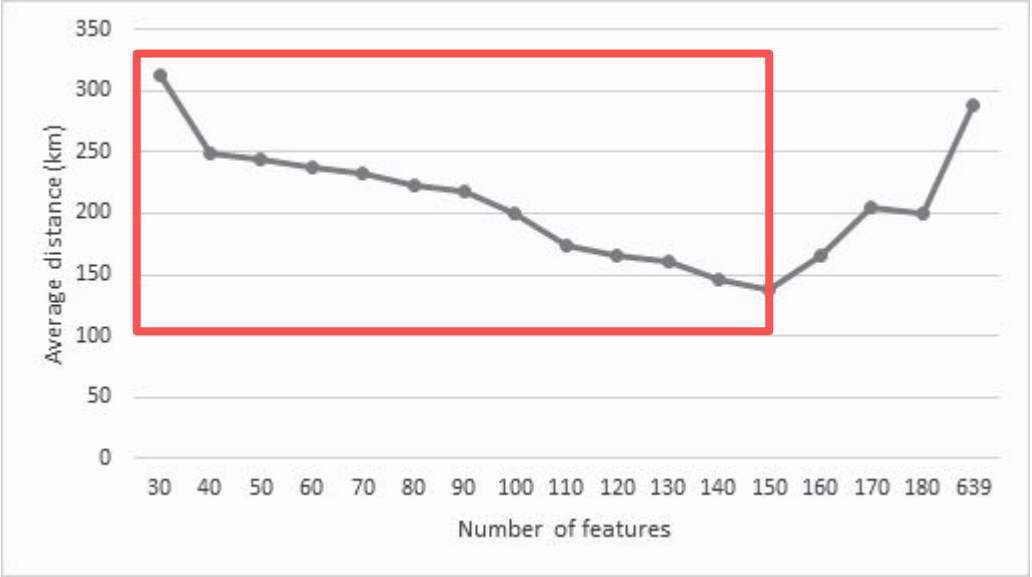
Recursive feature elimination

2. Some variants are hardly ever used or show no geographic variation at all. Discard them first.

Single-pass feature elimination based on χ^2 score

Remove variables that are least statistically dependent on area

Lowest average distance with 150 variants



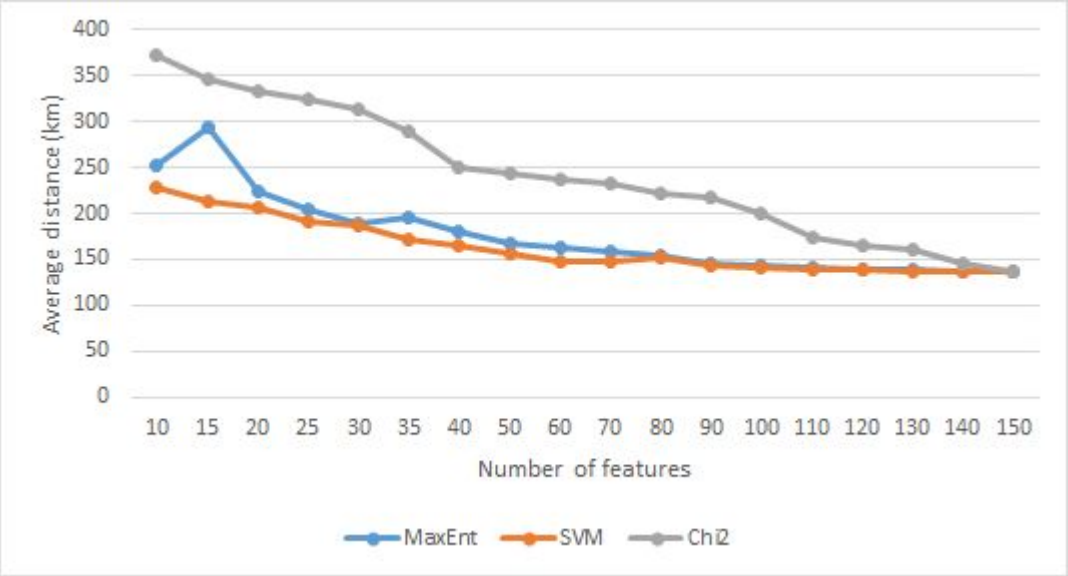
Recursive feature elimination

3. Train a classifier with the remaining variants, remove the one variant that contributes least to the classification, repeat (= recursive feature elimination).

We test two classifiers:
SVM and MaxEnt

Both classifiers achieve much better simulation results than the χ^2 method

MaxEnt slightly worse than SVM

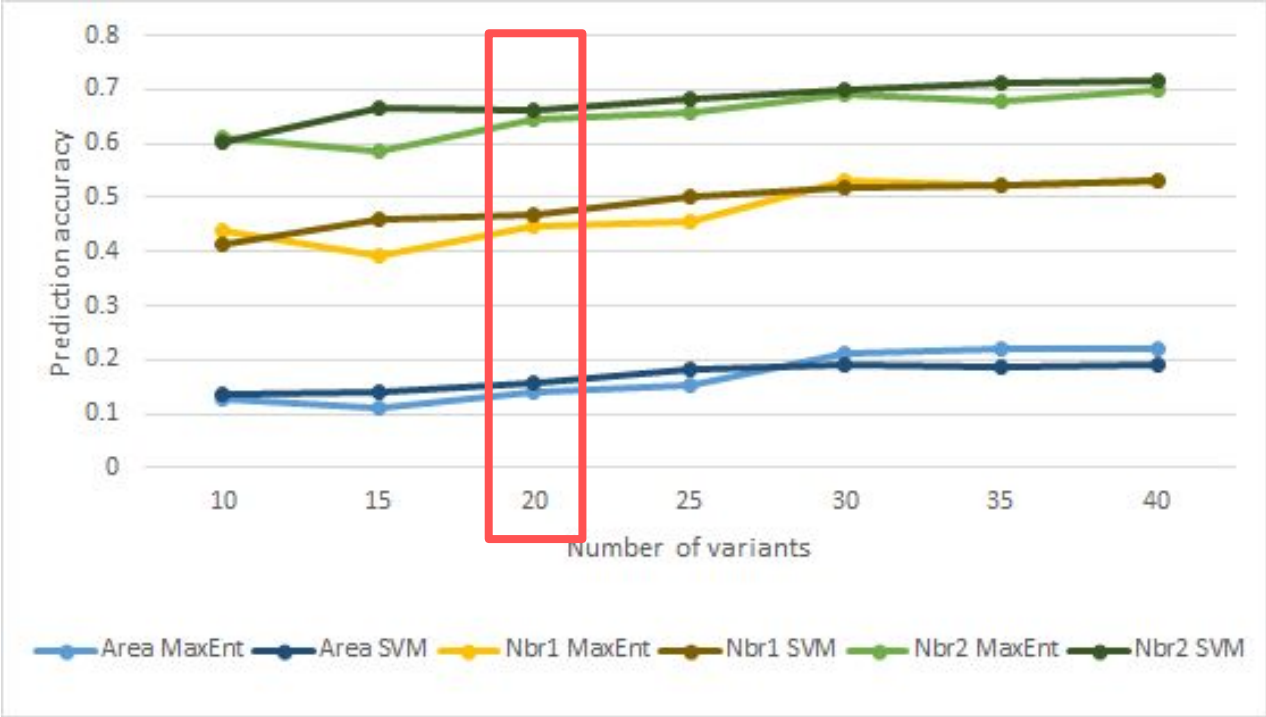


Recursive feature elimination

4. At the end, dynamically extend the areas to their immediate and second-order neighbors.

Simulation results with 20 variants / 17 questions:

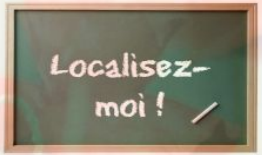
66.2% correct on second-order neighbors



Online speaker geolocalisation

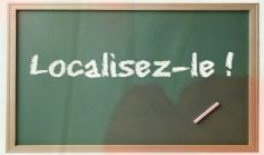
Donnez votre français à la science !

Connaissez-vous ces expressions de nos régions ? Participez au quiz !



Dites-nous comment vous parlez, on vous dira d'où vous venez !

Comment survivent, voyagent et meurent les particularismes linguistiques ? Répondez à quelques questions sur vos usages linguistiques.



Comment sont perçus les différents accents du français ? Essayez d'identifier la région d'origine des locuteurs que vous allez entendre.

Localisez-moi!

Question 2 sur 15

Comment appelez-vous ce fruit rouge, avec lequel on fait d'excellentes confitures ?

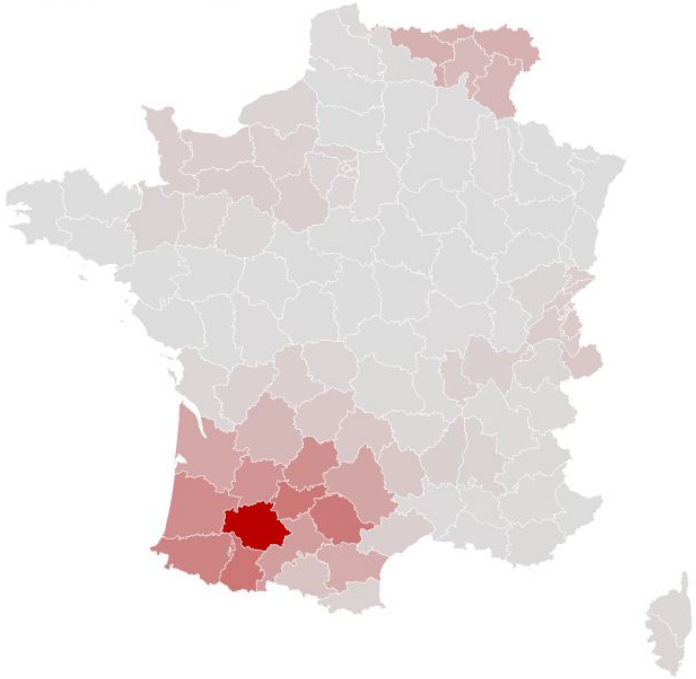
Myrtilles

Brimbelles

Localisez-moi!

Résultat: les départements en rouge représentent votre origine linguistique la plus probable.

Cliquez sur votre département d'origine



Aidez-nous à valoriser vos réponses en répondant à ce questionnaire

Où avez-vous passé la plus grande partie de votre jeunesse ?

Pays:

Code postal:

Adresse électronique (facultatif, ne sera pas diffusée à des tiers)

Année de naissance

Sexe

Sauvegarder les changements

Partagez sur Facebook

Online speaker geolocalisation

Three versions

- Feature elimination with MaxEnt 4000 participants
- Feature elimination with SVM 4000
- Manual selection of 15 questions 200

40% of participants provided sociolinguistic info (country+zip, age, gender, email)

Social networks sharing and media coverage

Online speaker geolocalisation

Crowdsourced data	Part	Best	5-Best	Neighb-1	Neighb-2
• Feature elimination ME	1631	11 %	43 %	40 %	62 %
• Feature elimination SVM	1679	13 %	47 %	47 %	64 %
• Manual selection	54	5 %	16 %	12 %	18 %
• Random		<1 %	4.5%	~4.5%	~9%

(110 areas - f-score)

Online speaker geolocalisation

Crowdsourced data	Part	Best	5-Best	Neighb-1	Neighb-2
• Feature elimination ME	1631	11 %	43 %	40 %	62 %
• Feature elimination SVM	1679	13 %	47 %	47 %	64 %
• Manual selection	54	5 %	16 %	12 %	18 %
• Random		<1 %	4.5%	~4.5%	~9%

Simulated data	Best	5-Best	Neighb-1	Neighb-2
• Feature elimination ME	14 %	49 %	47 %	64 %
• Feature elimination SVM	13 %	46 %	46 %	64 %
• Manual selection	10 %	36 %	40 %	57 %
			(110 areas - f-score)	

Discussion

- Attempt to apply machine learning techniques for question (and area) selection

  ⇒ estimate success of crowdsourced linguistic campaign before launch

- Automatic selection better than manual ? (to be confirmed)
- Crowdsourced geolocalisation also means data collection

⇒ donnezvotrefrancais.fr

Towards automatic geolocalisation of speakers of European French

Yves Scherrer & Jean-Philippe Goldman
University of Geneva

Recursive feature elimination

Retained features from the SVM classifier:

Pain au chocolat / chocolatine / couque au chocolat / ...
Ving[t]
Crayon de papier / de bois / gris / ...
Nonante / quatre-vingt-dix
Péguer
Gouttière / cheneau
Il est midi vingt / et vingt / vingt
Dîner / déjeuner
Pain aux raisins / escargot / schnäcke
Je vais y faire / le faire
Faire tomber / tomber / échapper
Séchoir / étendoir / étendage / tancarville
Moin[s]
Escargot / cagouille / luma
Dégun / personne

Retained features from the MaxEnt classifier:

Septante / soixante-dix
Ving(t)
Il est midi vingt / et vingt / vingt
Pain au chocolat / chocolatine / couque au chocolat / ...
Crayon de papier / de bois / gris / ...
Ça joue / ça va
Gorgée / schlouk / lichette
Gouttière / cheneau
Stan[d]
Empêtrer / encoubler / achouper / ..
Dîner / déjeuner
Péguer
Pain aux raisins / escargot / schnäcke
Séchoir / étendoir / étendage / tancarville
Papier ménage / Sopalin / essuie-tout

Si vous voulez parler d'une personne qui fumait et qui ne fu

Il a eu fumé (mais il ne fume plus).