

# Une thèse en TAL

Diversité linguistique, crowdsourcing et apprentissage automatique

---

Alice Millour

17 mars 2020

Atelier professionnel

# Plan

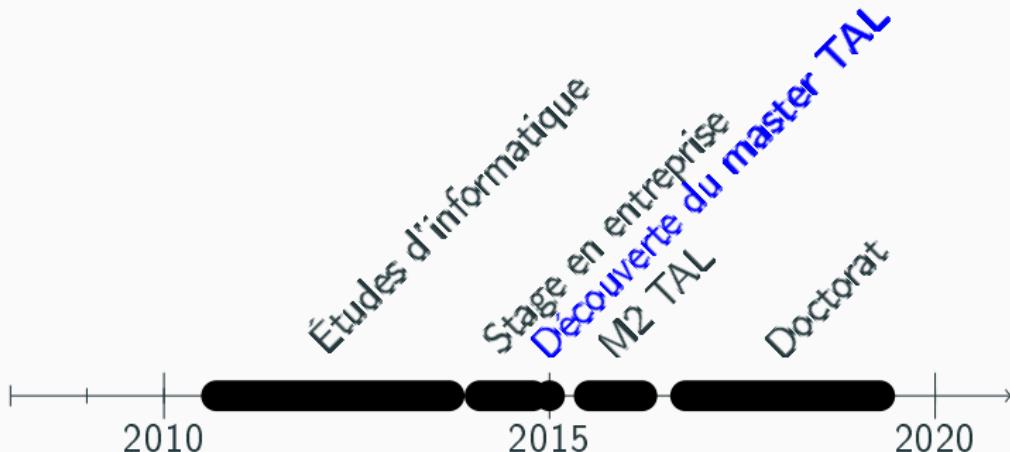
Une thèse en TAL, en quoi ça consiste ?

Diversité(s) linguistique(s)

Le crowdsourcing, solution à tous nos problèmes ?

Conclusions

# Mon parcours : Informatique + linguistique = TAL



Google

linguistique et informatique

Eviron 7 950 000 résultats (0,30 secondes)

Text Actualités Images Vidéos Maps Plus Paramètres Outils

Articles universitaires correspondant aux termes **linguistique et informatique**

... à partir d'une analyse linguistique de textes: réalisations ... - Jouria - Cite 48 fois  
Web Séminaire et Internat Linguistique - Amazigh - Cite 52 fois  
... le dialogue homme-machine. Méthodes Linguistiques et ... - UCLouvain - Cite 47 fois

Linguistique informatique — Wikipédia  
[https://fr.wikipedia.org/wiki/Linguistique\\_informatique](https://fr.wikipedia.org/wiki/Linguistique_informatique) •  
La Linguistique Informatique est un champ interdisciplinaire basé sur une modélisation symbolique  
(à base de règles) ou statistique du langage naturel étudié.  
Introduction Méthode Probabiliste Autre aux linguistes Linguistique de corpus

Spécialité Linguistique informatique (LI) - Université Paris Diderot  
<https://formation.univ-paris-diderot.fr.../specialite-linguistique-informatique-li> •  
Spécialité Linguistique informatique (LI) L'UFR de Linguistique de l'Université Paris-Diderot offre depuis plus de 20 ans un cursus complet de Linguistique

Master LPi - Langue et informatique (PI) - Offre de formation  
<http://sorbonne.fr.../master-lpi-langue-et-informatique-pi-programme-422.htm...> •  
Type de diplôme: Master; Domaine: Arts, Lettres, Langues; Mention: Philosophie, Linguistique; Spécialité: Langue et Informatique, Pratique du master ...

# Une thèse en TAL, en quoi ça consiste ?

Résoudre un problème

# Une thèse en TAL, en quoi ça consiste ?

## Résoudre un problème

- Lire
- Prendre des décisions
  - Comment découper mon problème ?
  - Quelle est la meilleure approche ?
- Faire des expériences
- Analyser les résultats
- Évaluer son approche

# Une thèse en TAL, en quoi ça consiste ?

## Résoudre un problème



## Faire connaître son travail

- Lire
- Prendre des décisions
  - Comment découper mon problème ?
  - Quelle est la meilleure approche ?
- Faire des expériences
- Analyser les résultats
- Évaluer son approche

# Une thèse en TAL, en quoi ça consiste ?

## Résoudre un problème

- Lire
- Prendre des décisions
  - Comment découper mon problème ?
  - Quelle est la meilleure approche ?
- Faire des expériences
- Analyser les résultats
- Évaluer son approche



## Faire connaître son travail

1. Écrire des articles (de recherche)
2. Soumettre son travail à la relecture d'autres chercheurs
3. Présenter son travail (posters, présentations orales etc.)



Intégrer les retours à son travail

# Une thèse en TAL, en quoi ça consiste ?

## Résoudre un problème

- Lire
- Prendre des décisions
  - Comment découper mon problème ?
  - Quelle est la meilleure approche ?
- Faire des expériences
- Analyser les résultats
- Évaluer son approche



⇒  
Sans oublier...

My thesis is written in



## Faire connaître son travail

1. Écrire des articles (de recherche)
2. Soumettre son travail à la relecture d'autres chercheurs
3. Présenter son travail (posters, présentations orales etc.)

Intégrer les retours à son travail

# Une thèse en TAL, en quoi ça consiste ?

## Et plus concrètement ?

“Similarités Textuelles Sémantiques Translingues : vers la Détection Automatique du Plagiat par Traduction”

“Modélisation du comportement des usagers à risque sur les réseaux sociaux”

“Vers des moteurs de recherche "intelligents" : un outil de détection automatique de thèmes. Méthode basée sur l'identification automatique des chaînes de référence”

“Traitement Automatique de la Langue Naturelle et interprétation : Contribution à l'élaboration d'un modèle informatique de la Sémantique Interprétabilité”

“Veille épidémiologique multilingue : une approche parcimonieuse au grain caractère fondée sur le genre textuel”

# Une thèse en TAL, en quoi ça consiste ?

## Et plus concrètement ?

“Similarités Textuelles Sémantiques Translingues : vers la Détection Automatique du Plagiat par Traduction”

“Modélisation du comportement des usagers à risque sur les réseaux sociaux”

“Vers des moteurs de recherche "intelligents" : un outil de détection automatique de thèmes. Méthode basée sur l'identification automatique des chaînes de référence”

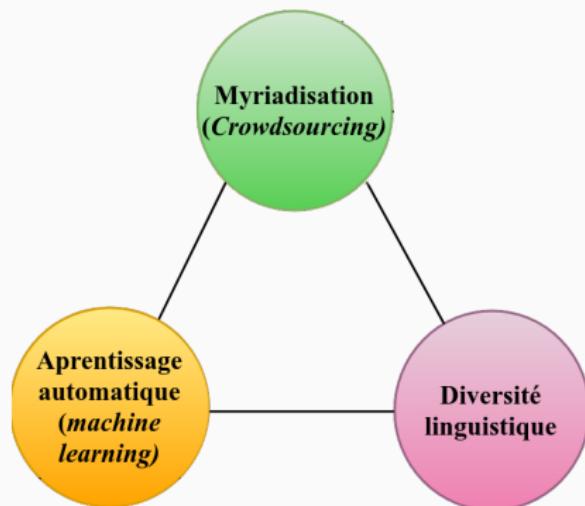
“Traitement Automatique de la Langue Naturelle et interprétation : Contribution à l'élaboration d'un modèle informatique de la Sémantique Interprétabilité”

“Veille épidémiologique multilingue : une approche parcimonieuse au grain caractère fondée sur le genre textuel”

“Construction de ressources langagières par myriadisation pour le traitement automatique des langues peu dotées : le cas des langues de France”

# Le sujet de ma thèse

Construction de ressources langagières par myriadisation  
(*crowdsourcing*) pour le traitement automatique des langues peu  
dotées : le cas des langues de France



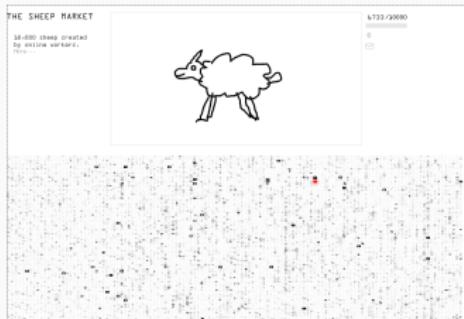
# Quelques définitions à avoir en tête

Myriadisation = crowdsourcing = production participative de données

*La myriadisation (crowdsourcing) consiste en un appel ouvert visant à faire produire des données (des entrées encyclopédiques, un dessin, un vote, etc.) à une masse de gens, aujourd'hui principalement via Internet.*



**WIKIPÉDIA**  
L'encyclopédie libre



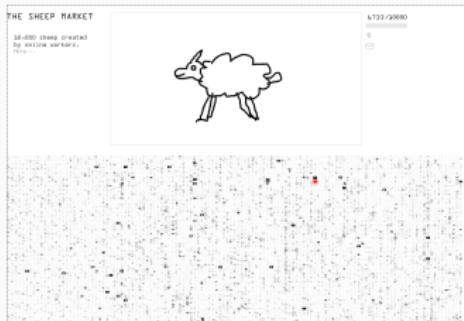
# Quelques définitions à avoir en tête

Myriadisation = crowdsourcing = production participative de données

*La myriadisation (crowdsourcing) consiste en un appel ouvert visant à faire produire des données (des entrées encyclopédiques, un dessin, un vote, etc.) à une masse de gens, aujourd'hui principalement via Internet.*



**WIKIPÉDIA**  
L'encyclopédie libre



⇒ appliqué aux ressources langagières (corpus bruts, annotés, lexiques etc.)

# L'apprentissage automatique

apprentissage automatique (*machine learning*)

Champ d'étude qui consiste à développer des méthodes visant à faire réaliser à une machine une tâche difficile ou complexe à décrire (grand nombre de paramètres).

- vision : reconnaissance d'objets, de visages
- recherche d'information
- analyse de données (Ex : financières, comportement d'utilisateurs etc.)
- systèmes de recommandations (Ex : YouTube, Netflix etc.)
- publicité ciblée
- locomotion de robots
- traitement automatique des langues
- etc.

## Quelques définitions à avoir en tête

apprentissage automatique (*machine learning*) pour le TAL

- classification de tweets selon leur polarité
- attribution de parties du discours (catégories grammaticales)
- regroupements de textes (langue, registre, thème)
- etc.

## Quelques définitions à avoir en tête

apprentissage automatique (*machine learning*) pour le TAL

- classification de tweets selon leur polarité
- attribution de parties du discours (catégories grammaticales)
- regroupements de textes (langue, registre, thème)
- etc.

⇒ fournir suffisamment de données (d'exemples) à un algorithme pour qu'il puisse “apprendre” à réaliser la tâche souhaitée

## Quelques définitions à avoir en tête

langues peu dotées

immense majorité des langues n'appartenant pas au groupe  
restreint des langues « bien dotées »

# Quelques définitions à avoir en tête

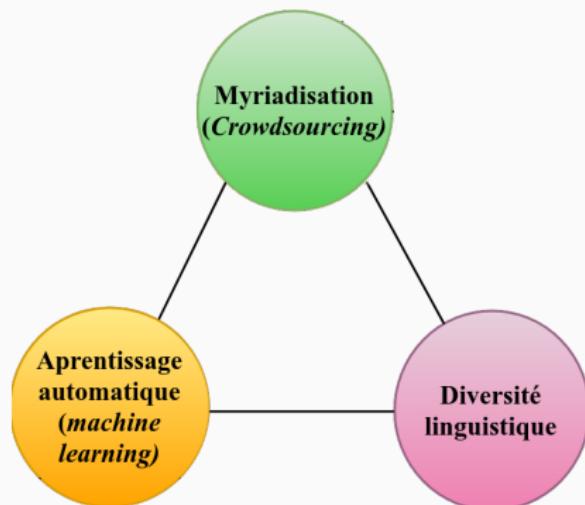
## langues peu dotées

immense majorité des langues n'appartenant pas au groupe  
restreint des langues « bien dotées »

- dictionnaire en ligne
- moteur de recherche
- outil de traduction automatique
- correcteur orthographique
- etc.

# Le sujet de ma thèse

Construction de ressources langagières par myriadisation  
(*crowdsourcing*) pour le traitement automatique des langues peu  
dotées : le cas des langues de France



# Plan

1. Une thèse en TAL, en quoi ça consiste ?

2. Diversité(s) linguistique(s)

Diversité linguistique réelle

Diversité linguistique dans les TICs

Diversité linguistique et TAL

3. Le crowdsourcing, solution à tous nos problèmes ?

4. Conclusions

# La diversité linguistique ?

- 24 langues parlées par plus de 50 millions de locuteurs
- 137 langues parlées par plus de 5 millions de locuteurs
- plusieurs milliers au total (7 097 ?)

Nom	Autres noms	Famille	ISO 639-3
aari	?a:ri (autonyme)	langues afro-asiatiques	aiw
abaknon	inabaknon, capuleño	langues austronésiennes	abx
abau		langues papoues	aau
abaza		langues caucasiennes	abq
abé	abbé, abbey, abi	langues nigéro-congolaises	aba
abénaqui		langues amérindiennes	aaq, abe
abidji		langues nigéro-congolaises	abi
abinomn		langues papoues	bsa
abipón		langues amérindiennes	axb
abkhaze		langues caucasiennes	abk
abom		langues papoues	aob
abon		langues nigéro-congolaises	abo
abouré	abure, abule, akaplass, abonwa	langues nigéro-congolaises	abu
abron	brong, bron, doma, gyaman	langues nigéro-congolaises	abr

Sources : <http://ethnologue.com> et [https://fr.wikipedia.org/wiki/Liste\\_de\\_langues](https://fr.wikipedia.org/wiki/Liste_de_langues)

# Plan

1. Une thèse en TAL, en quoi ça consiste ?

2. Diversité(s) linguistique(s)

Diversité linguistique réelle

Diversité linguistique dans les TICs

Diversité linguistique et TAL

3. Le crowdsourcing, solution à tous nos problèmes ?

4. Conclusions

# Les technologies de l'information et de la communication

**Tout ce qui permet aux personnes et aux organisations  
d'interagir avec le monde numérique**

terminaux (ordinateurs, téléphones, tablettes), accès à internet,  
logiciels, applications, données, transactions etc.

# Les technologies de l'information et de la communication

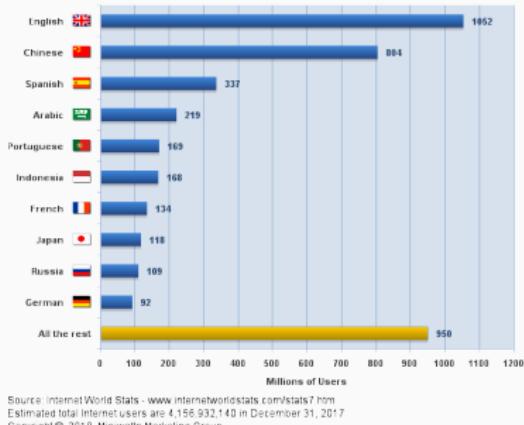
Tout ce qui permet aux personnes et aux organisations  
d'interagir avec le monde numérique

terminaux (ordinateurs, téléphones, tablettes), accès à internet,  
logiciels, applications, données, transactions etc.  
... et toutes nouvelles formes de productions langagières



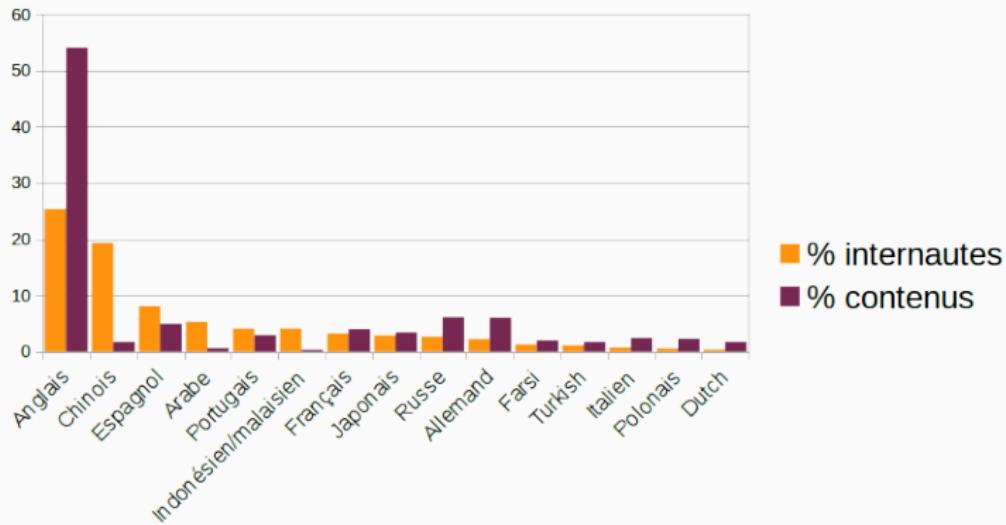
# Diversité linguistique sur Internet

- “pénétration d'internet” : 54,5 % (4,2Md internautes)
- langues du top 10 : 77,2 % des internautes



- 200 langues représentées dans le “top 10 millions de sites internet” (dont 160 comptent pour moins de 0.1%)
- taux de croissance nb. internautes (2000-2018) : +1 052,2 %  
Top 10 : +1 091,9 % - autres langues : +935,8 %

# Répartition des contenus



Source : <https://w3techs.com>

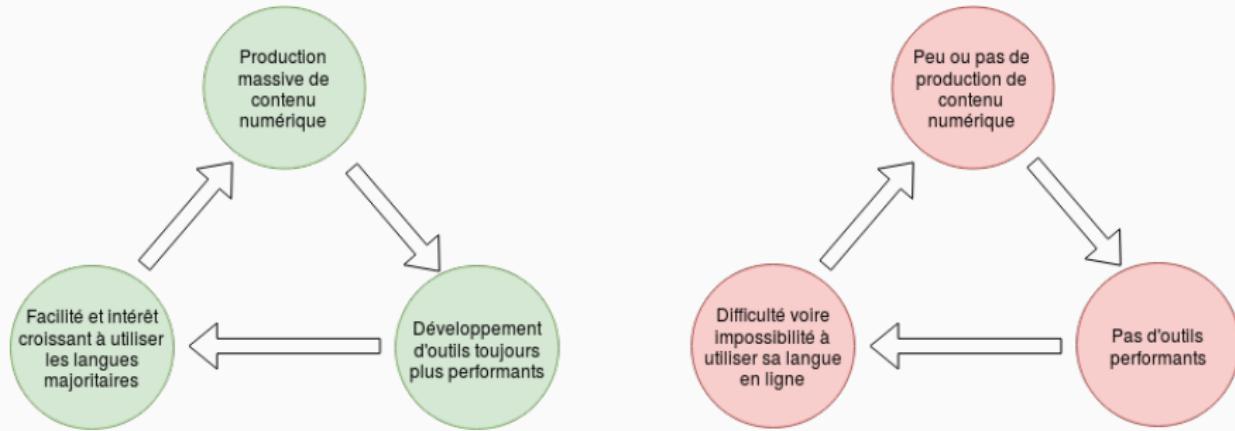
Internet n'est pas représentatif des internautes

Question n°1

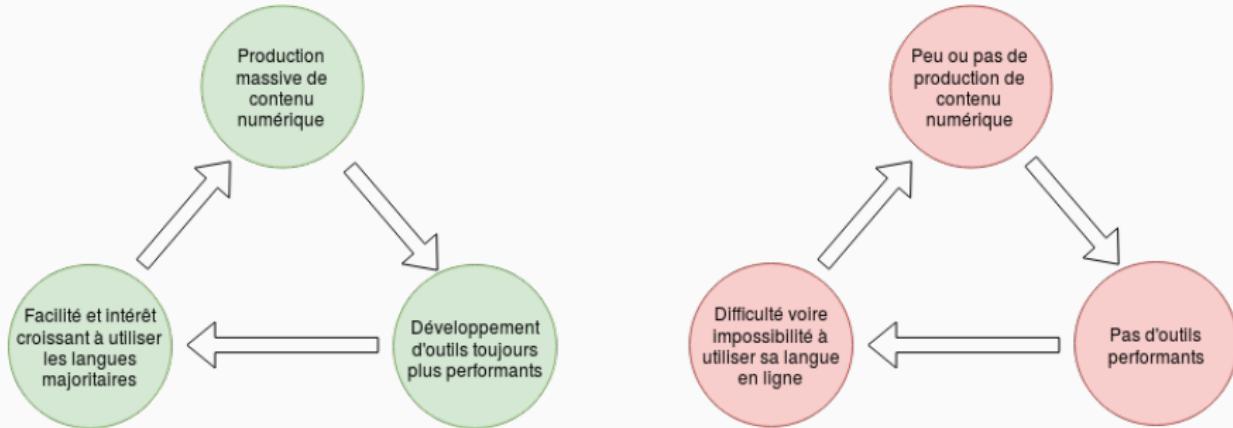
Est-ce que c'est grave ?



# Les cercles vertueux et vicieux de la présence en ligne



# Les cercles vertueux et vicieux de la présence en ligne



- ⇒ les internautes privilégient les langues majoritaires en ligne
- ⇒ accès inéquitable à l'information pour les internautes
- ⇒ certaines langues ne passent pas le “cap” de l'informatisation

# Les facteurs de disparition des langues

Une langue est menacée si elle...

- n'est plus en expansion
- perd ses fonctions de communication dans la vie sociale
- est parlée par moins de 100 000 locuteurs
- n'est plus pratiquée dans un ou plusieurs domaines de la vie quotidienne

Sources : *Halte à la mort des langues*, Claude Hagège / Projet UNESCO : *Atlas des langues en danger dans le monde*

# Les facteurs de disparition des langues

Une langue est menacée si elle...

- n'est plus en expansion
- perd ses fonctions de communication dans la vie sociale
- n'est pas parlée par plus de 100 000 locuteurs
- n'est plus pratiquée dans un des domaines de la vie quotidienne (école, travail, médecine, culture etc.)

Sources : *Halte à la mort des langues*, Claude Hagège / Projet UNESCO : Atlas des langues en danger dans le monde

# Internet n'est pas représentatif des internautes

Question n°1

Est-ce que c'est grave ?

Ça dépend de l'intérêt qu'on porte à la diversité linguistique

# Faut-il défendre la diversité linguistique ?



# Faut-il défendre la diversité linguistique ?

Pour

Contre

# Faut-il défendre la diversité linguistique ?

## Pour

- sauvegarde du patrimoine culturel, des croyances, des savoir-faire etc.
- garantit la diversité de **pensée**



Sources : lejournal.cnrs.fr, *La pensée unique et la diversité des langues*, Claude Hagège

## Contre

- risque de replis identitaires
- sauvegarde de langues pas pratiques, peu “performantes” etc.
- obstacle à l’intercompréhension universelle



## Le mythe de la Tour de Babel

*« Toute la terre avait une seule langue et les mêmes mots. [...] Allons ! descendons, et là confondons leur langage, afin qu'ils n'entendent plus la langue, les uns des autres. Et l'Éternel les dispersa loin de là sur la face de toute la terre et leur donna tous un langage différent. »*

## Le mythe de la Tour de Babel

« *Toute la terre avait une seule langue et les mêmes mots. [...] Allons ! descendons, et là confondons leur langage, afin qu'ils n'entendent plus la langue, les uns des autres. Et l'Éternel les dispersa loin de là sur la face de toute la terre et leur donna tous un langage différent.* »

⇒ [Le plurilinguisme](#)

(plus de la moitié de la population mondiale est au moins bilingue)

source : [courrierinternational](#)

# Faut-il défendre la diversité linguistique ?

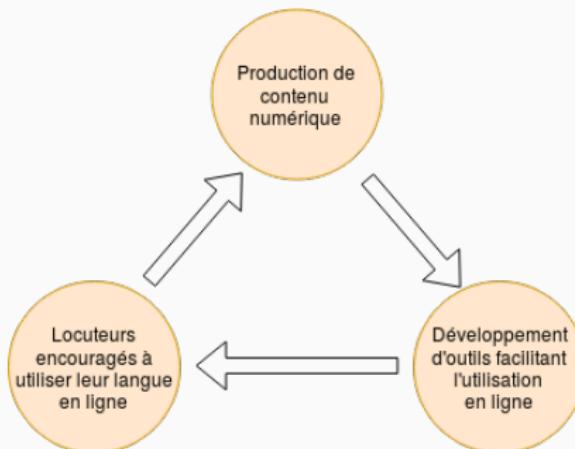
- Internet n'est pas représentatif des internautes
  - Question n°1 : est-ce que c'est grave ?
  - Ça dépend de l'intérêt qu'on porte à la diversité linguistique
- Question n°2 : faut-il défendre la diversité linguistique ?

Pour échapper à l'uniformisation linguistique, oui

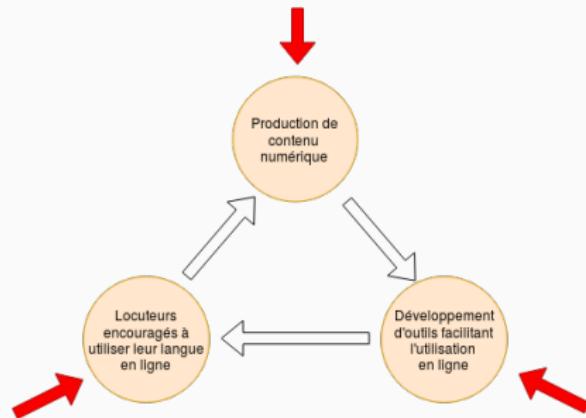
Question n°3

Comment faire ?

# Comment améliorer cette diversité ?



# Comment améliorer cette diversité ?



La recherche en TAL a un rôle à jouer

# Plan

1. Une thèse en TAL, en quoi ça consiste ?

2. Diversité(s) linguistique(s)

Diversité linguistique réelle

Diversité linguistique dans les TICs

Diversité linguistique et TAL

3. Le crowdsourcing, solution à tous nos problèmes ?

4. Conclusions

## Diversité linguistique et TAL

Comment mesurer la diversité linguistique en TAL ?

# Diversité linguistique et TAL

## Comment mesurer la diversité linguistique en TAL ? Google Translate :

Détecter la langue	Esperanto	Khmer	Russe
Afrikaans	Estonian	Kirghiz	Samoan
Albanais	Finnish	Kurdî	Serbe
Allemand	French	Laothian	Sesotho
Amharique	Frison	Latin	Shona
Anglais	Gaelique (Écosse)	Letton	Sindhi
Anzbe	Galician	Lithuanian	Slovaque
Armenien	Gallows	Luxembourgish	Slovene
Azérl	Géorgien	Macedonian	Somali
✓ Basque	Grec	Malaisien	Sundanais
Bengali	Gujarati	Malayalam	Sukhoi
Bélorusse	Hausa	Malgache	Svahili
Birmâne	Hawaiian	Malais	Tadjik
Bosniaque	Hébreu	Maori	Tagalog
Bulgare	Hindi	Marathi	Tamoul
Catalan	Hmong	Mongol	Tchèque
Cebuano	Hongrois	Nederlandse	Telugu
Chichewa	Igbo	Népalais	Thaï
Chinois	Indonésien	Norvégien	Turc
Cingalais	Irlandais	Ouzbek	Ukrainien
Corée	Islandais	Pêchi	Uru
Corse	Italien	Punjabi	Vietnamien
Créole haïtien	Japonais	Persan	Xhosa
Croate	Javaans	Polones	Yiddish
Danois	Kannada	Portugais	Yoruba
Espagnol	Kazakh	Roumain	Zoulou

100 langues

# Diversité linguistique et TAL

Piove i gatti ?

Français▼



Italien▼



Il pleut des cordes

Piove i gatti

# Diversité linguistique et TAL

Piove i gatti ?

Français▼



Italien▼



Il pleut des cordes

Piove i gatti

Français▼



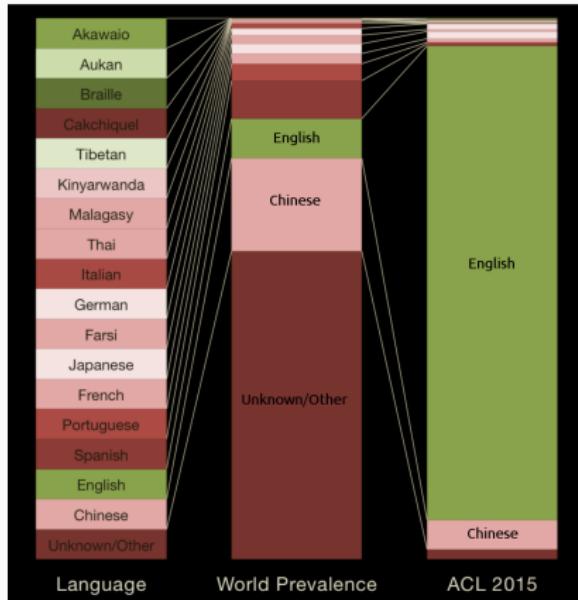
Anglais▼



Il pleut des cordes.

Raining cats and dogs.

# Diversité linguistique et TAL



ACL 2015 (source : <http://www.junglelightspeed.com/languages-at-acl-this-year/>)

**Écrasante majorité des recherches en TAL concernent l'anglais**

## Le problème de la diversité linguistique en TAL

- peu de chercheurs/de linguistes maîtrisent les langues concernées
- peu valorisant pour les chercheurs d'obtenir des résultats moyens sur des langues peu populaires
- **les ressources linguistiques nécessaires coûtent cher et sont longues à produire**

## Question n°4

Alors que fait-on en pratique ?

## Question n°4

Alors que fait-on en pratique ?



Question n°4

Alors que fait-on en pratique ?



## Question n°4

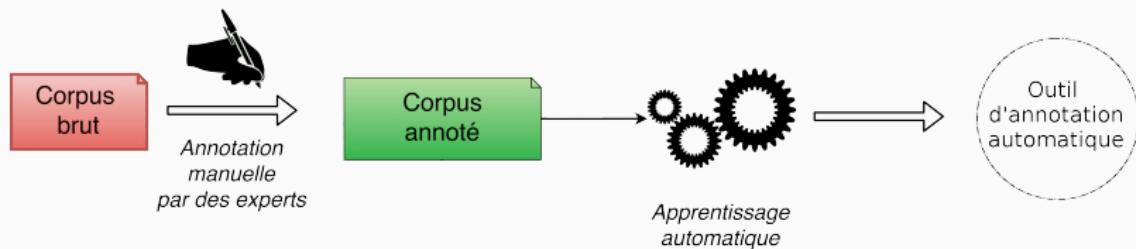
Alors que fait-on en pratique ?



# Plan

1. Une thèse en TAL, en quoi ça consiste ?
2. Diversité(s) linguistique(s)
3. Le crowdsourcing, solution à tous nos problèmes ?
  - L'idée générale
  - Exemple d'application
4. Conclusions

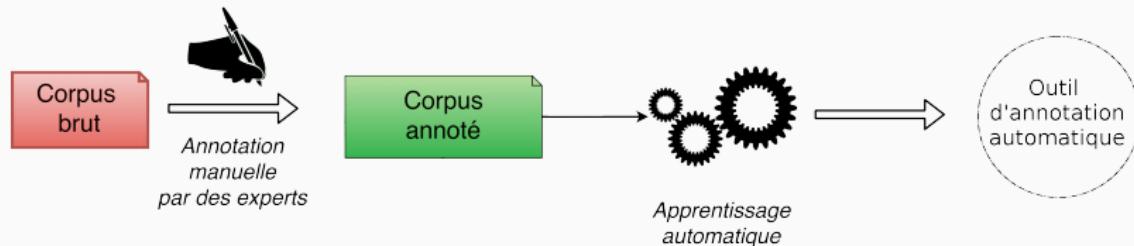
# Le crowdsourcing en TAL



**L'objectif :** faire “annoter” des locuteurs pour obtenir des exemples afin d’entraîner les algorithmes de classification

- Associer le bon “sentiment” à un tweet
- Associer la bonne catégorie grammaticale à un mot (annotation morphosyntaxique)
- etc.

# Le crowdsourcing en TAL



## Comment faire ?



# Plan

1. Une thèse en TAL, en quoi ça consiste ?
2. Diversité(s) linguistique(s)
3. Le crowdsourcing, solution à tous nos problèmes ?

L'idée générale

Exemple d'application

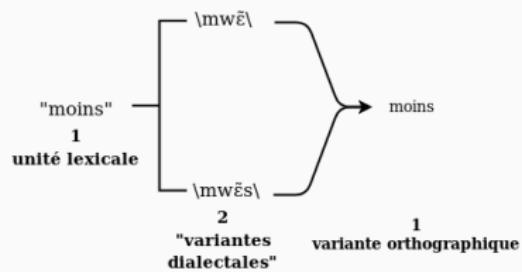
4. Conclusions

## cadre de ma thèse

- l'alsacien (200 000 locuteurs, 6 à 8 variantes, pas de norme orthographique consensuelle)

# La problématique de la variabilité dans les langues non standardisées

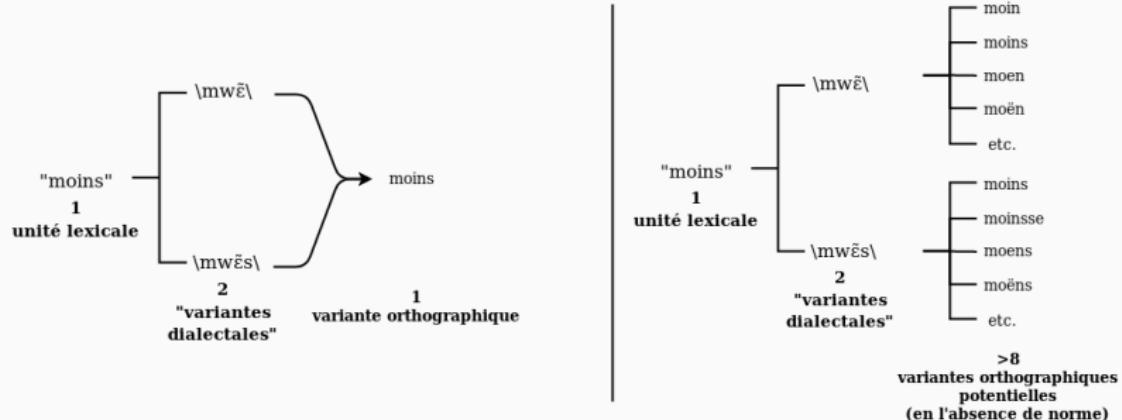
Variation phonologique avec “ Moins ”



en présence d'une norme  
orthographique

# La problématique de la variabilité dans les langues non standardisées

## Variation phonologique avec “ Moins ”

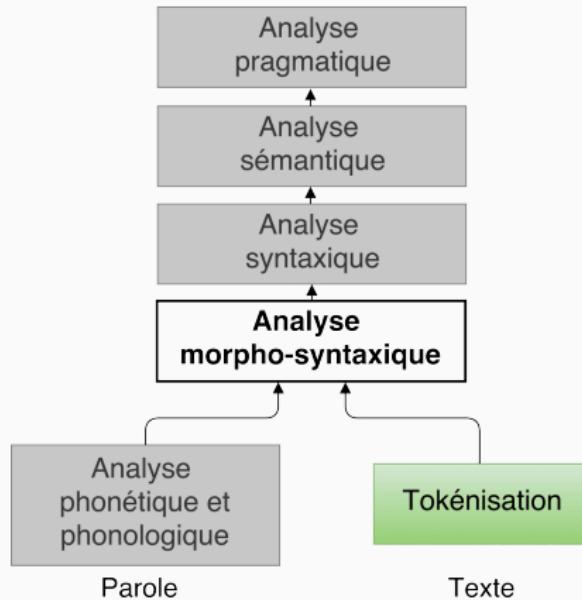


en présence d'une norme  
orthographique

en l'absence de norme  
orthographique

# cadre de ma thèse

- l'alsacien (200 000 locuteurs, 6 à 8 variantes, pas de norme orthographique consensuelle)
- l'annotation en parties du discours



# Exemple de plateforme : Recettes de Grammaire

The screenshot shows the homepage of the 'Recettes de Grammaire' platform. At the top, there are two boxes showing global statistics: one for the main account (135 participants, 6 recipes, 112 annotated words, 7 proposed alternative words) and one for the user 'milbmann' (52 points, 0 recipes, 38 annotated words, 6 proposed alternative words). The main header 'Recettes de Grammaire' is displayed over a background image of vegetables. Below the header, a sub-header reads 'Construisons ensemble des ressources linguistiques pour l'alsacien !'. The navigation bar includes links for Accueil, Gérer, Recettes, Contribuer, Trouver une recette, Contact, and Help. The main content area features several sections: 'Recette du jour' (Kugelhopf), 'Aujourd'hui, je contribue!', 'Classements', 'Recettes à valider', and buttons for Nouvelle recette, Annoter des recettes, and J'ajoute des variantes.

Statistiques globales

- 135 participants
- 6 recettes
- 112 mots annotés
- 7 mots alternatifs proposés

Mes statistiques

- 52 points
- 0 recettes
- 38 mots annotés
- 6 mots alternatifs proposés

Recette du jour

Kugelhopf

Recette de : milbmann

Aujourd'hui, je contribue !

Recettes de Grammaire est une plateforme collaborative qui recueille :

- Des recettes de cuisine (of Elsässisch !)
- Des annotations grammaticales servant à développer de nouveaux outils pour le traitement automatique de l'alsacien
- Des variantes orthographiques ou dialectales permettant un meilleur traitement de la variation en alsacien

Nouvelle recette

Annoter des recettes

J'ajoute des variantes

Je partage une recette

J'aide la science grâce à mes connaissances

J'aurais dit ça autrement

Classements

Recettes Annotations Variantes

1. milbmann (73 annotations)

- 2. Géry (58 annotations)
- 3. Gamer (38 annotations)
- 4. Maelle (38 annotations)
- 5. recettes\_Olca (34 annotations)

Recettes à valider

Aucune recette

Voir toutes les recettes

<http://bisame.paris-sorbonne.fr/recettes>  
[Millour and Fort, 2019]

Et ça marche ?

## Résultats à ce jour

Première expérience (annotation pure) :

53 participants

26 234 annotations produites en 3 mois

un corpus annoté de 8310 mots distincts

Seconde expérience (recettes) :

9 recettes (corpus de 2 057 mots)

619 annotations (350 mots)

122 mots alternatifs

**Des données de qualité en (trop) faible  
quantité**

# Plan

Une thèse en TAL, en quoi ça consiste ?

Diversité(s) linguistique(s)

Le crowdsourcing, solution à tous nos problèmes ?

Conclusions

# Conclusions

- la diversité linguistique passe par les technologies du langage, et le **TAL a un rôle à jouer**
- le crowdsourcing permet de recueillir des ressources uniques et de qualité
- même si les participants sont *a priori* motivés, les plateformes sont loin d'être autonomes

# Conclusions

- la diversité linguistique passe par les technologies du langage, et le **TAL a un rôle à jouer**
- le crowdsourcing permet de recueillir des ressources uniques et de qualité
- même si les participants sont *a priori* motivés, les plateformes sont loin d'être autonomes

**sensibiliser les locuteurs** de langues en danger de l'importance de la présence de leur langue en ligne

# Conclusions

- la diversité linguistique passe par les technologies du langage, et le **TAL a un rôle à jouer**
- le crowdsourcing permet de recueillir des ressources uniques et de qualité
- même si les participants sont *a priori* motivés, les plateformes sont loin d'être autonomes

**sensibiliser les locuteurs** de langues en danger de l'importance de la présence de leur langue en ligne

tâcher de **répondre aux besoins des locuteurs** autant qu'il répondent à nos besoins en ressources

## Bonus

Quest Game : "Katana and Grand Guru : A Game of the Lost Words"

(avec Marianne Araneta, Karën Fort, Ivana Lazić Konjik, Yann-Alan Pilatte, et Annalisa Raffone)



# Des questions ?

A circular diagram illustrating various Natural Language Processing (NLP) concepts and their interrelationships. The central concept is "machine learning". Other concepts are arranged around it, connected by arrows indicating their relationship to machine learning:

- open source
- gamification
- corpus
- développement web
- traitement des données
- diversité linguistique
- morphosyntaxe
- langues peu dotées
- crowdsourcing
- game with a purpose
- éthique
- optimisation de processus
- expérience utilisateur
- lexiques
- active learning
- annotation

*Merci !*

-  Millour, A. and Fort, K. (2019).  
**À l'écoute des locuteurs : production participative de ressources langagières pour des langues non standardisées.**  
In *Revue TAL : numéro spécial sur les langues peu dotées (59-3)*. Association pour le Traitement Automatique des Langues.