

Donnez votre Français à la Science ! Internet et la documentation de la diversité linguistique : présentation de la plateforme et premiers résultats

Julie Glikman^{1,*}, *Christophe Benzitoun*², *Jean-Philippe Goldman*³, *Yves Scherrer*⁴, *Mathieu Avanzi*⁵, et *Philippe Boula de Mareuil*⁶

¹ LiLpa, Université de Strasbourg, France

² ATILF, Université de Lorraine, France

³ Laboratoire d'Analyse et Technologie du Langage, Université de Genève, Suisse

⁴ Department of Digital Humanities, University of Helsinki, Finlande

⁵ FNRS & Centre VALIBEL, Université catholique de Louvain, Belgique

⁶ LIMSI, CNRS & Université Paris-Saclay, Orsay, France

Résumé. La plateforme de production participative (*crowdsourcing*) *Donnez votre Français à la Science* (DFS) vise à collecter des données linguistiques en mettant l'accent sur la variation en français européen. Les données recueillies dans le cadre d'un projet de production participative sont utiles pour une meilleure connaissance des différents usages de la langue, mais cela permet également de tisser des liens entre communauté scientifique et grand public en rendant ces derniers acteurs de la recherche et en les encourageant à suivre les enquêtes à venir. Les deux principales activités décrites ici sont : 1. une étude linguistique sur la connaissance de la variation lexicale régionale avec une rétroaction immédiate et 2. un système de géolocalisation des locuteurs, à savoir un quiz dont l'objectif est de deviner l'origine géographique du participant sur la base de la comparaison de ses réponses aux données précédemment recueillies dans d'autres enquêtes. Les évolutions de la plateforme, en cours de développement, sont également évoquées en conclusion.

* Contact : glikman@unistra.fr

Abstract. *Donnez votre français à la science ! (Give your French to Science!)* Crowdsourcing and linguistic variation: presentation and first results. The crowdsourcing platform called *Donnez Votre Français à la Science* (DFS, i.e. “Give your French to Science”) aims at linguistic data collection and language documentation, with a special focus on regional variation in European French. Not only is the gathered data useful for scientific studies, but providing feedback to the general public is also important in order to reward participants, encourage them to follow upcoming surveys, and foster exchanges with the scientific community. The two main activities described here are 1. a linguistic survey on lexical variation with immediate feedback and 2. a speaker geolocalization system, i.e. a quiz that guesses the linguistic origin of the participant on the basis of his/her answers compared to previously gathered linguistic data.

1 Introduction

Des enquêtes linguistiques peuvent de nos jours être aisément mises en place via le web et les smartphones grâce à ce que l’on appelle la méthode de production participative (*crowdsourcing* en anglais), parfois encore appelée *myriadisation* (Fort 2017). Cette méthode est utilisée dans différents domaines, par exemple en psychologie¹, en études de marketing² ou encore pour la description de données archéologiques³. Elle permet désormais, en linguistique, de collecter de grandes quantités de données en peu de temps (Cook et al. 2013). Elle peut ainsi aider à décrire les phénomènes linguistiques qui sont sous-représentés, voire totalement absents des corpus traditionnels (formes régionales et non-normatives, entre autres). On peut, par exemple, envisager d'utiliser le crowdsourcing de manière longitudinale afin de comparer des instantanés successifs de variables linguistiques et ainsi décrire la variation diachronique, en republiant la même enquête à plusieurs années d'intervalle par exemple, ou en comparant sur une même enquête les participants par tranches d'âge.

Au cours des dernières années, divers projets de crowdsourcing ont été mis en place pour la collecte de données linguistiques grâce à des applications web, comme *Le Français de nos régions* (Avanzi et al. 2016) pour les aires francophones de l'Europe et du Canada, *L'Atlas der deutschen Alltagssprache* (Möller et Elspaß 2015) pour la variation régionale de l'allemand, l'enquête *Harvard Dialect* (Vaux et Bert 2013) pour la variation régionale aux Etats-Unis, *Verba Alpina* (Krefeld et Lücke 2014) pour les dialectes parlés dans les Alpes, et plus récemment, les sites jumeaux

tonaccent / dindialäkt, mettant l'accent sur la perception de l'accent suisse en français et dans les dialectes suisses allemands (Goldman et al. 2018a).

D'autres actions ont été menées telles que des jeux de géolocalisation, visant à prédire la provenance de l'utilisateur sur la base d'un ensemble d'objets qui devaient être nommés. Alors que certains projets ont été distribués sous forme d'applications pour smartphone dans le but de documenter la variation dialectale en anglais et en allemand (Leemann et al. 2016), d'autres (principalement en dehors de la communauté scientifique) ont été conçus comme des sites Web à visée ludique, par exemple les quiz du journal belge *Le Soir* (« Quel français de Belgique parlez-vous ?⁴ ») ou de la Télévision Suisse Romande (avec « le Parlomètre romand⁵ »).

Toutes ces initiatives fondées sur un thème populaire tel que la variation linguistique ont rencontré un vif succès auprès du grand public et ont aidé à recueillir un grand nombre de données linguistiques, permettant des études scientifiques novatrices dans ce domaine, ainsi que la publication d'ouvrages à destination du grand public (Avanzi 2017).

L'utilisation de la production participative s'est également développée pour l'analyse des données linguistiques, comme la production d'annotations linguistiques (Fort 2017, Millour et al. 2017, Munro et al. 2010). Des plateformes ont été créées à cet usage, telles que *l'Amazon's Mechanical Turk*, proposant des rémunérations à bas prix pour des tâches d'analyse parcellisées (voir Callison-Burch et Dredze 2010, et Sagot et al. 2011 pour une critique de l'approche).

Dans cet article, nous présentons la plateforme de production participative *Donnez votre français à la science !* (désormais DFS)⁶, dont l'objectif est de servir au recueil de données tout en fournissant des retours vulgarisés au grand public. Le retour au grand public est un aspect important de notre approche, pour le rendre acteur de la recherche et lui montrer concrètement ce que sa participation permet de construire, permettant ainsi de l'impliquer et de lui donner envie de participer à d'autres enquêtes. La principale ambition de la plateforme DFS est de devenir une plateforme de référence dans laquelle les participants intéressés pourront être sollicités régulièrement, mais aussi d'être utilisable par tout linguiste pour créer des enquêtes en ligne et profiter du panel existant de participants. L'objectif de DFS est ainsi de proposer une alternative libre, gratuite et spécifiquement dédiée à la recherche en linguistique des sites générateurs d'enquêtes déjà existants, mais souvent limités, inadaptés ou payants (Google Forms⁷, LimeSurvey⁸, Qualtrics⁹...) Dans sa version actuelle¹⁰, la plateforme DFS comprend plusieurs types d'activités liées à la variation linguistique. Elle se limite pour l'instant à la langue française utilisée sur le continent européen (à savoir la Belgique, la France et la Suisse). Après une brève présentation de la

plateforme, nous exposons les premiers résultats de deux activités : un quiz linguistique avec une rétroaction immédiate vers le participant, et une activité de géolocalisation. Nous évoquons enfin en conclusion les évolutions de la plateforme en cours de développement.

2. Présentation de la plateforme DFS

La plateforme DFS consiste en un site internet collaboratif destiné à toute personne (chercheurs, étudiants, enseignants, public non-spécialiste, etc.) s'intéressant à la diversité du français au sens large. Cette plateforme a pour mission de servir de portail pour le recueil de données linguistiques au moyen de la méthode dite de la « production participative » (angl. *crowdsourcing*), tout en fournissant un retour au grand public. Différents aspects de la variation du français peuvent ainsi être étudiés : outre la variation diatopique, la plateforme permettra également de s'intéresser à d'autres variétés, comme le « parler jeune », le parler des banlieues, le parler des apprenants, à travers l'étude de leurs caractéristiques propres. Le retour informatif et ludique au grand public permet de montrer directement aux participants leur contribution à la recherche scientifique. Nous pensons que cela les incitera à poursuivre leur participation aux autres enquêtes ainsi qu'à partager leur expérience avec leur entourage pour attirer de nouveaux participants.

La version actuelle de la plateforme DFS est basée sur l'outil de crowdsourcing PyBossa¹¹, qui permet de développer diverses activités en parallèle. Notre objectif est de recueillir des données pour les linguistes et de fournir une rétroaction aux participants pour certaines activités. À la fin d'une tâche, le participant est invité à partager ses résultats sur les réseaux sociaux (cf. parties 3 et 4 ci-dessous), contribuant ainsi à populariser la plateforme. Parmi les activités proposées dans sa version actuelle, nous présentons dans cet article les premiers résultats de la mise en place d'un quiz linguistique (partie 3) et d'une tâche de géolocalisation (partie 4). Ces deux activités donnent une rétroaction directe aux participants. De tels retours vers l'utilisateur sont très importants à nos yeux, car ils favorisent la compréhension du projet de recherche et ainsi les échanges entre la communauté scientifique et le grand public. Ils permettent en outre de fidéliser les participants en vue de futures enquêtes.

3. Quiz linguistique

En utilisant les données des enquêtes précédentes sur la variation du français européen (Avanzi et al. 2016), nous avons créé un quiz dans lequel les participants sont invités à tester leur connaissance du français régional. Le

quiz prend la forme d'un questionnaire à choix multiples où une seule réponse est correcte. Le participant est immédiatement informé de l'exactitude de sa réponse et une brève explication linguistique est donnée. À la fin du quiz, le participant se voit attribuer un score final, qui peut être partagé sur les réseaux sociaux. C'est également à la fin du quiz que l'on propose aux participants, de manière optionnelle, de remplir un court formulaire de renseignements (pays et code postal de la localité où ils ont passé la plus grande partie de leur enfance¹², sexe, année de naissance) pour permettre de valoriser leurs réponses. Ce quiz possède différents objectifs. Par son aspect ludique (il s'agit d'un jeu, avec un score final, et les différentes réponses proposées sont souvent exprimées d'une manière humoristique), il est censé attirer des participants et leur donner envie d'utiliser la plateforme de manière générale, et de participer ainsi aux autres activités, plus spécifiquement orientées vers le recueil de données (type formulaires d'enquêtes linguistiques, également accessibles via la plateforme). À travers les explications données au cours du quiz, nous informons les participants sur les résultats récents de la recherche en linguistique. Enfin, à travers les scores obtenus par les participants, le linguiste dispose d'un indicateur de la vitalité et de la connaissance passive des régionalismes au sein de la population, qu'on peut comparer aux résultats des précédentes enquêtes. Les analyses présentées ci-dessous portent sur les résultats pour l'activité *Quiz des expressions de nos régions* entre le 10 mars et le 6 octobre 2017, pour les 2879 participants ayant répondu à l'ensemble des 12 questions que comporte le quiz¹³, puis plus spécifiquement pour les 1077 participants (soit seulement 37% des 2879) ayant rempli le questionnaire sur leur origine.

Dans les enquêtes précédentes (Avanzi et al. 2016), les participants devaient indiquer quel(s) mot(s) ou quelle(s) expression(s) ils utiliseraient en contexte quotidien et familier, à partir d'une liste à choix multiples. Ces enquêtes, de type déclaratif, ont permis de dresser des cartes linguistiques pour la variation régionale en français (voir Avanzi 2017, et le blog *français de nos régions*). Dans le quiz, ce n'est plus l'utilisation, mais la connaissance, éventuellement passive, de ces mots qui est testée, le but étant en même temps de privilégier l'aspect ludique et informatif (réponse correcte immédiatement donnée, avec une explication linguistique et visualisation de la zone d'emploi du terme sur une carte, score final, avec possibilité de partager son score sur les réseaux sociaux). Les résultats aux questions du quiz, couplés avec les indications géographiques des participants ayant complété le formulaire de renseignements, constituent des données analysables qui peuvent nous permettre de mieux connaître la diffusion des items interrogés. En effet, ces résultats permettent de construire des

visualisations de zones de connaissance des items du quiz, qui peuvent ainsi être comparées aux cartes des premières enquêtes.

Les formulations des questions du quiz peuvent être réparties en deux types. Une première catégorie correspond à des définitions, du type « que signifie tel terme ? ». On peut considérer ainsi que ce type de question interroge la compréhension, même passive, du terme. C'est le cas pour des termes comme *chocolatine* (Question : *Qu'est-ce qu'une chocolatine ?* Réponses possibles : *Une tranche de pain recouverte de pâte à tartiner ; Un chocolat fourré à la praline ; Une pâtisserie de pâte feuilletée fourrée d'une barre de chocolat, aussi appelée "pain au chocolat" ; Une boisson au chocolat chaud aromatisée à la vanille*), *nareux* (Question : *Quand on dit que quelqu'un est nareux, cela veut dire que cette personne :* Réponses possibles : *est difficile avec la nourriture ; a un gros nez ; a le nez bouché ; a la nausée*), *dégun* (Question : *S'il y a dégun, cela veut dire que :* Réponses possibles : *Il n'y a personne ; Il y a urgence ; Il n'y a rien à faire ; Il y a eu des dégâts*), ou encore *péguer* (Question : *Que signifie le verbe péguer ?* Réponses possibles : *Puer ; Appuyer avec insistance ; Coller légèrement ; Bégayer*), *foehn*, *schluck* et *tancarville*.

La deuxième catégorie de questions porte plus explicitement sur la connaissance de la variation régionale, du type « comment appelle-t-on ceci à tel endroit », comme pour la *poche*, dénomination du sac plastique à Toulouse (Question : *Quel autre nom donne-t-on au sac plastique dans la région de Toulouse ?* Réponses possibles : *Un sachet ; Un cornet ; Une poche ; Un nylon*), ou encore la variation sémantique régionale, comme pour l'emploi différencié selon les pays francophones d'Europe du verbe *dîner* (Question : *Si un Suisse ou un Belge vous invite à dîner, cela veut dire qu'il vous invite :* Réponses possibles : *à manger à midi ; à manger le soir ; à partager une pâtisserie avec un thé aux alentours de 16h*). Outre la connaissance de la variation, la question sur la prononciation du chiffre *80* en Belgique (Question : *Comment dit-on 80 en Belgique ?* Réponses possibles : *Quatre-vingts ; Huitante ; Octante*) permet aussi de tester les représentations sur la langue, la plupart des participants de France ayant obtenu de mauvais scores à cette question.

Le score par question pour les 2879 participants montre de grands écarts selon les questions :

Tableau 1. Score de réponses correctes par question, trié par pourcentage de réponses justes.

Question	Score	Pourcentage
chocolatine	2669	92.13 %
sac plastique	2293	79.15 %
tancarville	2180	75.25 %
schluck	2015	69.55 %
diner	1891	65.27 %
dégun	1765	60.93 %
péguer	1679	57.96 %
torchon	1600	55.23 %
foehn	1217	42.01 %
nareux	1172	40.46 %
cayon	746	25.75 %
80	425	14.67 %

On peut noter l'écart entre la question sur la *chocolatine*, qui a obtenu 92,13% de bonnes réponses, indiquant une bonne compréhension du terme, et la question sur la prononciation de 80, qui en a obtenu seulement 14,67%. Ce score nous semble montrer la méconnaissance des participants de l'usage en Belgique. Cette méconnaissance est cependant associée à une représentation de la variation. En effet, la bonne réponse au quiz est *quatre-vingt*, bien que les Belges francophones utilisent les variantes *septante* et *nonante* (pour 70 et 90). On peut ainsi supposer de ces résultats que les participants ayant donné une réponse autre (*octante* ou *huitante*) devaient connaître cette variation, et l'étendre à 80¹⁴.

Pour les 1077 participants ayant renseigné leur origine, nous avons pu établir des cartes de diffusion des items selon les résultats aux questions¹⁵. On peut ainsi comparer l'aire d'utilisation de *chocolatine* (effectuée d'après l'origine des participants au quiz, sur données lissées, plus les réponses de la zone sont correctes, plus la couleur est foncée), par rapport à son aire d'emploi (effectuée d'après les résultats aux précédentes enquêtes, Avanzi et al. 2016) :

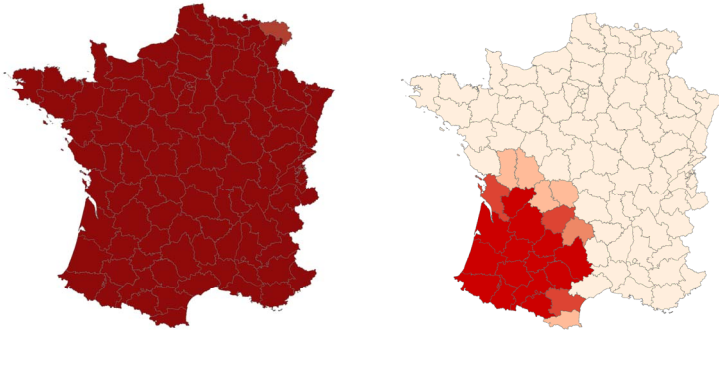


Fig. 1. Carte de réponses correctes au quiz vs carte d'emploi (enquêtes) de *chocolatine*.

Cette diffusion peut facilement s'expliquer par le développement de l'utilisation de ce terme dans la presse, notamment en lien avec les débats épilinguistiques qu'il a suscités¹⁶, comme on peut le voir dans le graphique d'évolution de *chocolatine* dans le corpus Europresse :

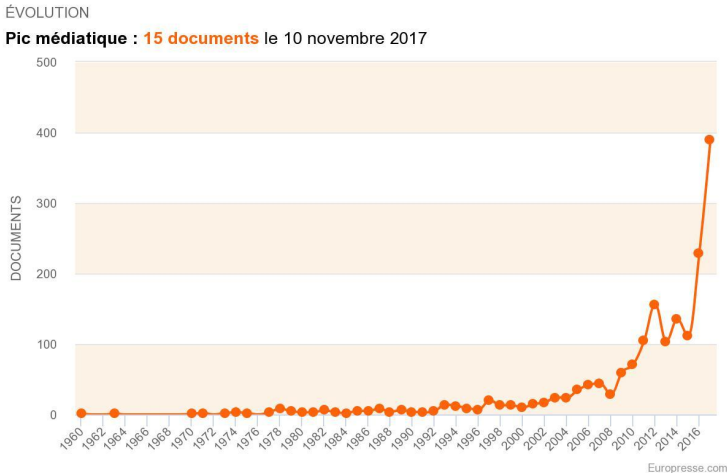


Fig. 2. Graphique Europresse d'évolution médiatique de *chocolatine*.

À l'inverse, on peut constater, pour d'autres items, que la zone de réponses correctes au quiz est quasiment identique à la zone de déclaration d'emploi du terme dans les enquêtes, ce qui peut être interprété comme un terme qui s'est peu diffusé hors de sa zone d'emploi, même en compréhension passive. C'est le cas pour l'adjectif *nareux* :

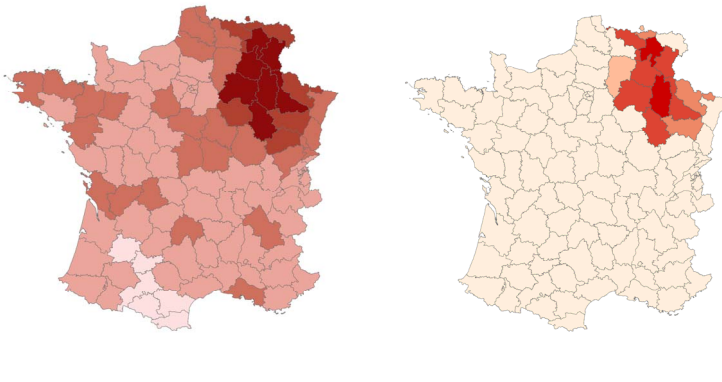


Fig. 3. Carte de réponses correctes au quiz vs carte d'emploi (enquêtes) de *nareux*.

Cette absence de diffusion du terme en dehors de sa zone d'emploi est également confirmée par sa faible représentativité dans la presse, comme le montre le graphique Europresse ci-dessous (le pic médiatique étant en outre principalement lié, là aussi, au discours épilinguistique, puisqu'il s'agit d'articles portant sur la variation en français régional, lors de la sortie du livre d'Avanzi (2017) :

ÉVOLUTION

Pic médiatique : 8 documents le 25 novembre 2017

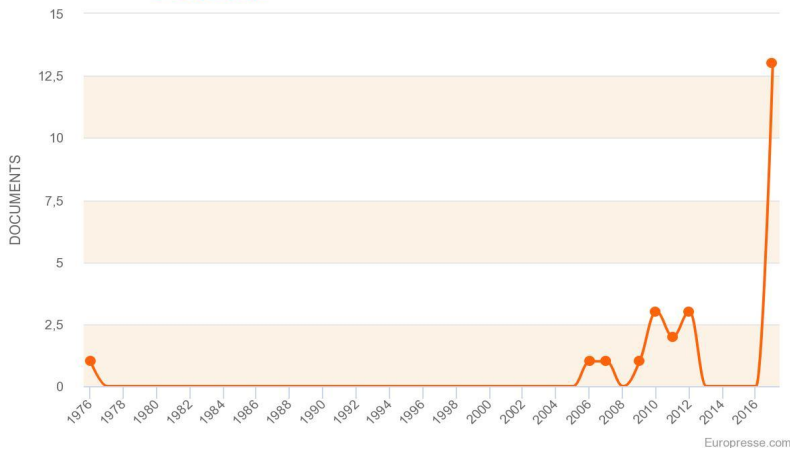


Fig. 4. Graphique Europresse d'évolution médiatique de *nareux*.

Cette différence est aussi observable dans le corpus Wortschatz¹⁷, qui range *chocolatine* et *chocolatines* dans les classes de fréquence 21 et 22

respectivement (36 et 14 occurrences), mais ne donne aucun résultat pour *nareux*, *nareuse*.

D'après nos analyses, les cartes de résultats au quiz pour les mots *cayon*, *péguer*, et *foehn* correspondent ainsi de manière assez proche aux cartes d'emploi des mots. Pour d'autres mots, la carte de résultats des quiz montre une diffusion qui s'étend hors de la zone d'emploi, tout en montrant une concentration de bonnes réponses plus élevée dans la zone d'emploi. C'est le cas pour *tancarville* (voir ci-dessous les cartes en fig. 5), *dégun*, ou *schluck*.

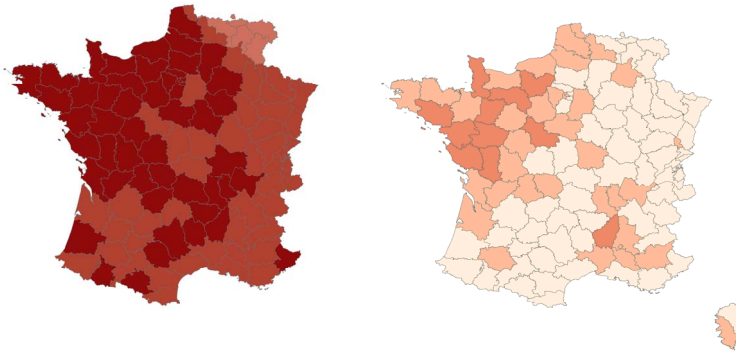


Fig. 5. Carte de réponses correctes au quiz vs carte d'emploi (enquêtes) de *tancarville*.

Enfin, dans le cas de la prononciation de 80, l'effet est inverse : en effet, les zones de réponses correctes sont principalement en Belgique francophone, alors que le terme effectivement utilisé en Belgique, en l'occurrence *quatre-vingt*, est utilisé dans l'ensemble de la France et de la Belgique :

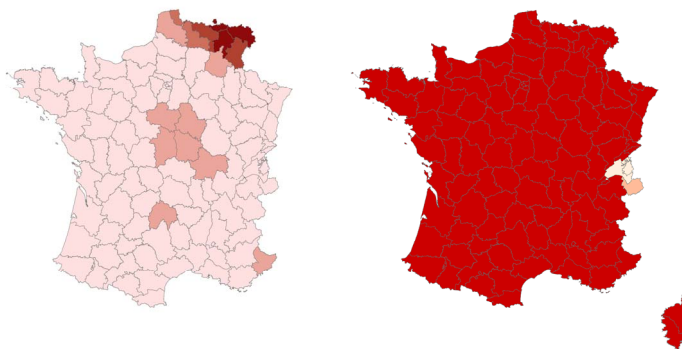


Fig. 6. Carte de réponses correctes au quiz vs carte d'emploi (enquêtes) de *quatre-vingt*.

Comme nous l'avons évoqué ci-dessus, ce résultat montre la représentation linguistique que les francophones de France ont sur le français de Belgique : ils sont au courant de la variation sur la dénomination des nombres, mais sans savoir l'appliquer dans son détail, et, par une sorte d'hypercorrection, ajoutent une spécificité régionale là où elle n'existe pas.

4. Géolocalisation

En utilisant en tant que matériel d'apprentissage les enquêtes déjà réalisées (Avanzi et al. 2016), nous avons également créé une activité centrée sur la géolocalisation. En plus de proposer une méthode de validation scientifique des données déjà recueillies, l'objectif de cette activité est de fournir une incitation ludique aux utilisateurs pour participer à d'autres tâches sur la plateforme DFS, et d'enrichir les données pour améliorer en permanence les activités.

À la suite de Leemann et al. (2016), nous définissons ici la géolocalisation comme la prédiction de la variété régionale d'un locuteur en lui posant une série de questions à propos de son usage linguistique. Compte tenu de nos motivations, la pertinence de la méthode de géolocalisation ne doit pas seulement être évaluée par le pourcentage de prédictions correctes, mais aussi par sa capacité à divertir et à faire réfléchir les participants potentiels. Trois paramètres peuvent avoir une influence sur le succès de la méthode : *N* - le nombre et le type de questions à poser (il ne faut pas poser plus de 20 questions, pour ne pas excéder la durée d'attention) ; *M* - le nombre et la taille des zones géographiques à prévoir (les zones doivent refléter de manière fidèle les aires de variation régionale du français, mais des zones trop grandes pourraient rendre le problème trivial et sans intérêt) ; *A* -

la précision des prévisions (la méthode doit évidemment faire d'aussi bonnes prédictions que possible, mais nous estimons qu'au moins 2/3 de prévisions correctes sont nécessaires pour un niveau d'implication durable des participants).

Dans les sections suivantes, nous donnons quelques détails sur la façon dont nous avons abordé ce problème d'optimisation (voir aussi Goldman et al. 2018b). En d'autres termes, nous voulons choisir, parmi les données linguistiques qui ont été recueillies précédemment, la meilleure série de questions (N étant aussi réduit que possible) avec le meilleur ensemble de zones de prédiction (les zones M étant aussi petites que possible) et la plus grande précision A. Afin d'essayer de prédire les chances de succès d'une tâche de géolocalisation avant son lancement, nous avons mis en place un cadre de simulation pour trouver les meilleurs réglages possibles.

4.1 Données

Les deux enquêtes de départ ont porté sur les régionalismes dans les aires francophones européennes (France, Belgique et Suisse) en 2015-2016 dans le cadre du projet *Français de nos régions* (Avanzi et al. 2016). Les questionnaires ont été remplis par 20.000 participants (tableau 2).

Tableau 2. Nombre de questions et de participants pour chaque enquête

Enquête 1	Enquête 2
Mai 2015- Mai 2016	Septembre 2015 - Mai 2016
40 questions	90 questions
12.000 participants	8.000 participants

Ces enquêtes abordaient principalement la question des variantes lexicales, et, dans une moindre mesure, les variantes (morpho-)syntaxiques et phonologiques, à travers des questions à choix multiples, illustrées par des photos. Cela pouvait prendre la forme de questions directes portant sur l'utilisation du mot (ex. *Utilisez-vous le mot s'entrucher?* Réponses possibles : *Oui - Non*) ou impliquer une définition d'un concept ou d'un objet (*Comment appelez-vous le morceau de tissu que l'on utilise pour lessiver les sols?* Réponses possibles : *Serpillière ; Wassingue ; Loque (à reloqueter) ; Panosse ; Torchon ; Pièce (à froter) ; Since ; Bâche ; Peille ; Toile*). Le nombre de réponses possibles varie de 2 à 10, et les réponses multiples étaient autorisées.

4.2 Expérience de simulation

Deux prétraitements importants ont été appliqués. D'une part, nous nous sommes arrêtés sur un ensemble de 109 zones administratives comme limite supérieure pour M : nous avons retenu 96 départements français, 7 cantons suisses (appartenant à la partie francophone de ce pays), et 8 provinces belges (appartenant à la partie francophone de ce pays). Bien que la plupart des participants à l'enquête aient fourni leur code postal, nous avons regroupé les locuteurs dans l'une de ces 109 zones pour éviter les problèmes de données insuffisantes dans les zones peu peuplées. D'autre part, nous avons apparié les participants de la première enquête avec les participants de la seconde enquête sur la base de leur origine géographique, conduisant à un ensemble de données de 6463 participants.

Afin d'évaluer les différents réglages des paramètres N, M et A, nous avons mis en place une expérience de simulation, en utilisant uniquement les données de l'enquête avec un algorithme d'exclusion (« leave-one-out »). L'idée de base consiste donc 1. à entraîner un modèle sur les données agrégées provenant de l'ensemble des participants sauf un, afin de prédire l'origine du participant exclu, et 2. à comparer la prédiction avec la réalité. Cependant, contrairement à un véritable algorithme d'exclusion, nous avons choisi de ne pas supprimer le participant test des données d'entraînement à des fins d'efficacité (en supprimant la nécessité d'élaborer un nouveau modèle pour chaque participant). Étant donné que les données d'entraînement sont agrégées et contiennent toujours plus d'un participant pour chaque point d'enquête et question, il n'y a jamais exactement le même point d'enquête dans le corpus d'entraînement et de test, ce qui nous permet d'utiliser ce raccourci méthodologique.

Nous avons envisagé deux approches pour trouver les meilleurs paramètres de géolocalisation, l'une fondée sur le partitionnement et la détection de schibboleth, et l'autre sur l'élimination de traits.

4.3 Partitionnement et détection de schibboleth

Notre approche comporte deux étapes : nous déterminons d'abord une partition d'aires optimale en utilisant la classification hiérarchique, et appliquons ensuite l'algorithme de détection de schibboleth de Prokić et al. (2012) pour trouver l'ensemble de questions le plus caractéristique pour chaque aire.

Les figures 7 et 8 illustrent trois solutions de classification hiérarchique utilisant différents algorithmes et différentes partitions cibles. Il convient de souligner que les données sources agrégées délimitent bien des régions

cohérentes tant d'un point de vue géographique que linguistique, ce qui suggère que la qualité des données d'enquête est bonne. Seule la méthode moyenne pondérée a montré des zones non-compactes au-dessus de 10 partitions.

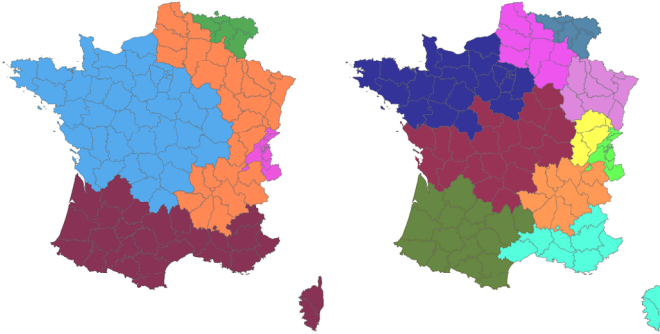


Fig. 7. Cartes obtenues en utilisant la méthode de Ward (5 partitions puis 10 partitions).

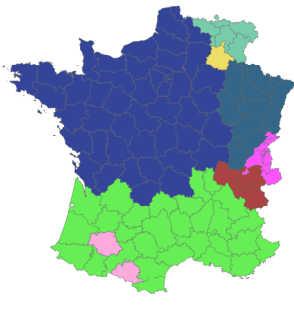


Fig. 8. Carte obtenue à l'aide de la méthode de la moyenne pondérée (10 partitions).

Nous avons choisi la méthode de Ward avec 10 partitions pour la suite de nos expériences. En ce qui concerne le choix optimal des questions, nous avons effectué plusieurs simulations de géolocalisation avec différents paramètres, dont trois sont illustrées ici :

1. 10 groupes de Ward et l'ensemble des 130 questions : précision de 65,1 %, mais les résultats sont très sensibles aux frontières de groupes : -24 % entre 4 et 5 groupes ; -21 % entre 10 et 11 groupes.

- Cela donne à penser qu'il est difficile de déterminer un nombre de groupes idéal et un algorithme optimal.
2. 10 groupes de Ward et 14 questions déterminées manuellement : précision de 67 %. Cela montre que quelques questions soigneusement sélectionnées donnent de meilleurs résultats que la prise en compte de l'ensemble des questions.
 3. 10 groupes de Ward et 20 questions déterminées par la détection de schibboleth : précision de 61,8%. Nous avons observé un choix de questions plutôt contre-intuitif, avec des variantes de français standard apparaissant dans la plupart des régions. Si cet algorithme donne comme variantes linguistiques les plus caractéristiques pour la Suisse francophone (zone verte sur la carte en haut à droite), les variantes *encoubler*, *septante*, *nonante*, *ça joue*, *souper*, la sélection est moins convaincante pour la Provence (zone de couleur cyan sur la carte en haut à droite) : *péguer*, *challer*, *soixante-dix*, *sèche-cheveux*, *quatre-vingt-dix*. Il faut également noter que les groupes sont définis sur l'ensemble des données, et non sur le sous-ensemble de questions résultant de la détection de schibboleth.

La combinaison du partitionnement hiérarchique et de la détection de schibboleth a donné des résultats mitigés, d'une part parce que le partitionnement est très sensible au choix de paramètres, et d'autre part parce que certaines des zones ne se définissent pas par des schibboleths clairs, mais plutôt par des légères différences dans la distribution des variantes. Nous nous sommes donc tournés vers une approche alternative.

4.4 Élimination de traits

Dans une deuxième approche, nous ne fixons pas les partitions à l'avance, mais gardons les 109 zones telles que définies ci-dessus et nous focalisons sur la recherche de l'ensemble optimal de questions. Pour cela, nous appliquons des techniques d'élimination de traits, comme détaillé ci-dessous. Une fois que l'ensemble optimal de questions a été déterminé, nous étendons dynamiquement les prévisions aux n meilleures zones ou voisins :

1. Étant donné que les variables linguistiques peuvent avoir plusieurs variantes avec des distributions différentes, nous avons traité chaque variante séparément et nous avons transformé les données de 130 n-aires à 639 variables binaires. Par exemple, nous convertissons la question 11-aire (avec 11 réponses possibles) *Comment appelez-vous le morceau de tissu que l'on utilise pour lessiver les sols ?*, en 11 questions binaires du type *Appelez-vous le morceau de tissu 'serpillière' ?*.

- Certaines variables ne sont presque jamais utilisées ou ne montrent aucune spécificité géographique. Nous les avons donc écartées en utilisant une procédure d'élimination de traits en une passe basée sur le χ^2 . En d'autres termes, nous avons supprimé les variables qui étaient statistiquement les moins dépendantes de la localisation. Cela élimine par exemple les variantes *chiffon*, *patte* et *lave-pont*, variantes très rarement sélectionnées par les participants de l'enquête. 150 variables correspondent à la distance moyenne la plus basse entre la prédiction et la réalité, comme le montre la figure 9. Nous avons pris en compte cette valeur pour les étapes suivantes.

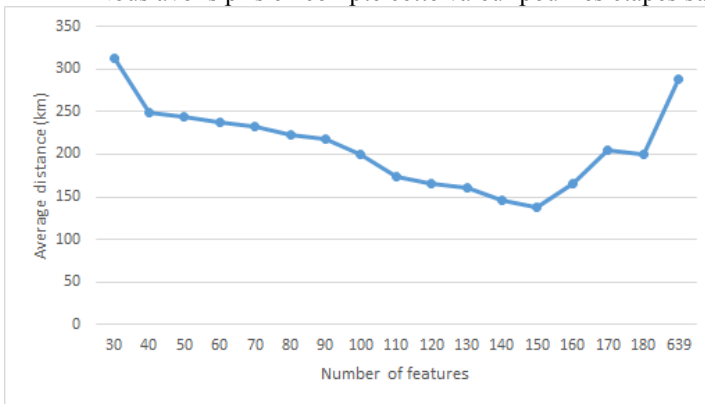


Fig. 9. Distance moyenne (en km) en fonction du nombre de variables

- Nous continuons à éliminer graduellement les variantes les moins pertinentes pour la géolocalisation, mais en utilisant un algorithme qui tient compte de l'interaction entre les variantes, la procédure récursive d'élimination de traits (Guyon et al. 2002). Cette procédure élimine de façon répétée la variante qui contribue le moins à la classification. Nous avons effectué deux expériences parallèles avec deux algorithmes de classification sous-jacents, SVM et MaxEnt. Les deux classifieurs obtiennent de bien meilleurs résultats de simulation que si nous avions continué à utiliser la méthode χ^2 . MaxEnt affiche une performance légèrement moins bonne que SVM, comme le montre la figure 10.

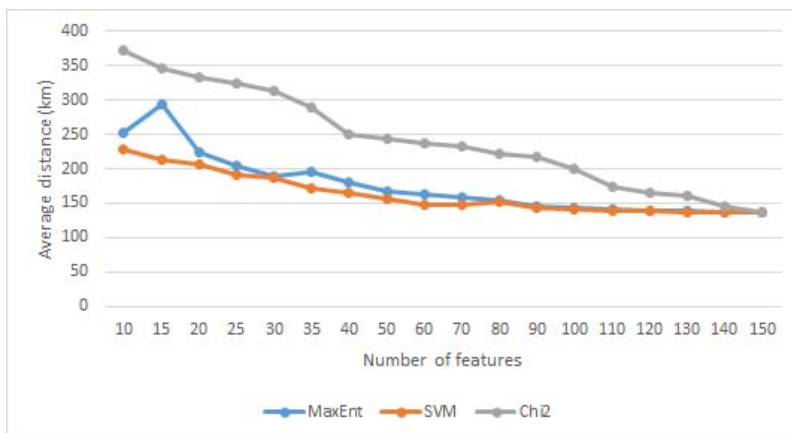


Fig. 10. Distance moyenne (en km) en fonction du nombre de variables retenues pour les classifieurs SVM, MaxEnt et χ^2 .

4. En fin de compte, nous avons utilisé les 109 zones et les distances calculées entre centroïdes. Nous avons également étendu dynamiquement les zones à leurs voisins immédiats et de second ordre.

La figure 11 montre les résultats de la simulation pour les deux classifieurs, compte tenu de la zone exacte, des voisins immédiats (Nbr-1) et des voisins de second ordre (Nbr-2). Avec 20 variables (représentant 17 questions n-aires), le score de précision était de 66,2 % sur les voisins de second ordre.

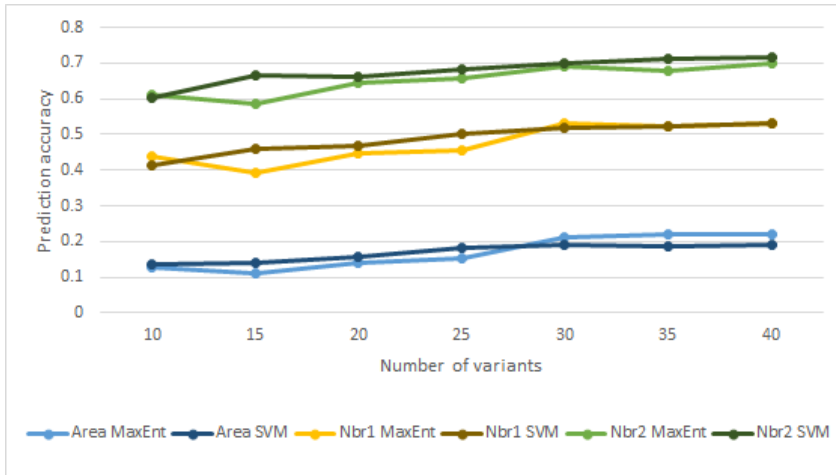



Fig. 11. Précision de la prédiction de 10 à 40 variables avec des classifieurs SVM et MaxEnt et compte tenu de la correspondance de zone exacte, les zones voisines immédiates (Nbr-1) et voisins de second ordre.


À partir de ces simulations, nous pouvons tirer quelques conclusions : premièrement, les résultats se stabilisent à environ 40 variables binaires, à savoir environ 30 questions n-aires. En second lieu, l'extension des prédictions aux voisins de premier ordre améliore la précision de 30% et l'extension aux voisins de second ordre l'améliore de 20% supplémentaires.

4.5 Mise en œuvre et résultats

L'enquête a été implémentée dans la plateforme DFS de façon similaire à la première activité, en utilisant les 20 questions n-aires résultant des approches d'élimination de traits SVM et MaxEnt. À la fin de l'activité, l'outil montre à chaque participant une carte représentant les aires francophones européennes avec une estimation statistique de son origine linguistique. Nous demandons également des informations sociolinguistiques aux participants (pays et code postal, âge, sexe) ainsi que leur adresse mail. Environ 40% d'entre eux ont fourni ces données.

Ci-dessous une capture d'écran du formulaire de saisie :

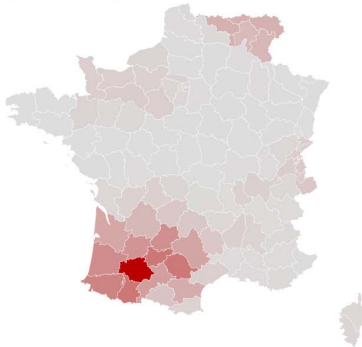

 Suggestions Crédits À propos Blog [↗](#) Contact [↗](#)



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:

Grâce au mécanisme de partage sur les réseaux sociaux et à la couverture médiatique de notre projet, 8000 personnes ont participé à l'expérience. Leur ont été appliqués alternativement les algorithmes de MaxEnt et SVM. Plus tard, nous avons élaboré une troisième enquête basée sur une sélection manuelle de 15 questions, qui a été proposée à 200 participants. Le tableau 3 montre les résultats de la classification pour les participants qui ont donné leur véritable localisation. Avec les deux méthodes automatiques, on atteint le seuil de précision souhaitée avec des tailles de surface et un nombre de variables comparables (environ 20). Cependant, les variables sélectionnées par la seconde approche correspondent intuitivement mieux aux variations observées dans les données de l'enquête initiale.

Tableau 3. Données du crowdsourcing (les pourcentages correspondent au f-score pour 109 zones)

	Part	Best	5-Best	Neighb-1	Neighb-2
MaxEnt	1631	11 %	43 %	40 %	62 %
SVM	1679	13 %	47 %	47 %	64 %
Sélection manuelle	54	5 %	16 %	12 %	18 %
Aléatoire		<1%	4.5 %	4.5%	9%

Notre tentative d'appliquer des techniques d'apprentissage automatique pour la sélection de questions (et de régions) conduit à une précision de 66 %,

comme mentionné ci-dessus. L'avantage principal de cette approche automatique est d'estimer les chances de succès de la campagne du crowdsourcing avant le lancement d'études plus fouillées.

5. Conclusion

La flexibilité de la plateforme DFS permet ainsi de créer différentes activités autour de la variation linguistique. Outre le recueil de données à travers les enquêtes linguistiques, elle permet la mise en place d'autres activités, à la fois ludiques et à visée informative. Ces activités constituent un retour vers le grand public, un moyen de l'intéresser à la diversité linguistique, et de lui donner envie de contribuer à la recherche, en *donnant son français* à son tour. Elles fournissent également des données linguistiques analysables, qui permettent de compléter les données existantes et contribuent ainsi à la documentation de la variation linguistique, comme nous l'avons vu pour le quiz et l'activité de géolocalisation. En effet, on a pu reprocher aux enquêtes linguistiques de type déclaratif de ne pas être forcément représentatives de l'usage réel des sujets parlants. Les associer à d'autres types d'activités, comme ici la connaissance, éventuellement passive, des items, ou la géolocalisation, permet de vérifier et compléter ces enquêtes.

Dans les développements futurs, nous visons la création facilitée de nouvelles enquêtes linguistiques par la communauté scientifique, à travers la mise en place d'un générateur de formulaires, et l'intégration d'un volet enregistrement de productions orales intégrées à une enquête linguistique. L'intégration d'un volet enregistrement via la méthode participative est en soi une innovation pour le français, et aucun site de formulaires ne fournit cette fonctionnalité actuellement. L'enregistrement vocal pourra servir à des expériences de lecture, mais aussi de recueil de productions orales non lues de mot simple, à partir d'images, ou encore de productions orales plus longues comme un récit (de type procédural, comme des descriptions de trajets ou de recettes, ou encore à partir d'images). Ces nouvelles données pourront être croisées aux données d'autres enquêtes et contribuer à la documentation fine de la variation linguistique. Conçue au début pour le français, la plateforme pourra à terme être ouverte à d'autres langues.

La nouvelle plateforme DFS comprendra ainsi diverses activités pouvant être avec ou sans rétroaction immédiate (comme le quiz ou la géolocalisation). Le générateur de formulaire permettra de construire son activité à partir de différents types de questions en entrée (comme image, texte ou son) et d'avoir différents types de réponses possibles en sortie (réponses à choix multiple, enregistrement vocal, ou encore réponse écrite). La plateforme comprendra également un espace blog pour le retour vers le

grand public des résultats d'enquêtes, et nous envisageons la mise à dispositions des données recueillies via la plateforme.

Remerciements

Cette recherche est financée par la DGLFLF (Délégation générale à la langue française et aux langues de France) et soutenue par ces partenaires académiques : Université de Strasbourg, Université de Genève, Université d'Helsinki, UC Louvain, LIMSI, laboratoire ATILF et Ortolang.

Références bibliographiques

- Avanzi, M., C. Barbet, J. Glikman, J. Peuvergne (2016). Présentation d'une enquête pour l'étude des régionalismes du français. *Actes du 5ème congrès mondial de linguistique française (CMLF), SHS Web of Conferences (Vol. 27, p. 03001). EDP Sciences.*
- Avanzi, M., Thibault, A. (à par.). Histoire, aréologie et valeur sociale du système vigésimal et de sa contrepartie décimale (70-80-90) en français. *Langages.*
- Avanzi, M. (2017). *Atlas du français de nos régions.* Paris : Armand Colin.
- Callison-Burch, C., M. Dredze (2010). Creating speech and language data with Amazon's Mechanical Turk. *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 1-12.
- Cook, M., J. Barker & Lecumberri, M. L. G. (2013). Crowdsourcing in Speech Perception. In Eskénazi, M., Levow, G.A., Meng, H., Parent, G. & Suendermann, D. (eds), *Crowdsourcing for Speech Processing : Applications to Data Collection, Transcription and Assessment.* Hoboken : John Wiley & Sons, 137-172.
- Fort, K. (2017). Experts ou (foule de) non-experts ? la question de l'expertise des annotateurs vue de la myriadisation (crowdsourcing). *Corela* [En ligne], HS-21 | 2017, mis en ligne le 20 février 2017, consulté le 21 février 2017. URL : <http://corela.revues.org/4835>
- Goldman, J.-P., S. Clematide, M. Avanzi, R. Tandler (2018a). Strategies and Challenges for Crowdsourcing Regional Dialect Perception Data for Swiss German and Swiss French. *LREC*, Miyasaki, Japon (à paraître).
- Goldman, J.-P., Y. Scherrer, J. Glikman, M. Avanzi, C. Benzitoun, P. Boula de Mareuil (2018b). Crowdsourcing Regional Variables and Automatic Geolocalisation of Speakers of European French. *LREC*, Miyasaki, Japon (à paraître).
- Guyon, I., J. Weston, S. Barnhill, V. Vapnik (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3) : 389-422.
- Krefeld, T., S. Lücke (2014). Verba Alpina - Der alpine Kulturraum im Spiegel seiner Mehrsprachigkeit. *Ladinia* 38 : 189-211.
- Leemann, A., M.-J. Kolly, R. Purves, D. Britain, E. Glaser (2016). Crowdsourcing language change with smartphone applications. *PLOS ONE.*
- Millour, A., K. Fort, D. Bernhard, L. Steible (2017). Vers une solution légère de production de données pour le TAL : création d'un tagger de l'alsacien par

- crowdsourcing bénévole. *Traitement Automatique des Langues Naturelles (TALN)*, 2017, Orléans, France. <hal-01516226v2>
- Möller, R., S. Elspaß (2015). Atlas zur deutschen Alltagssprache. Kehrein, R., A. Lameli, S. Rabanus (eds.) *Regionale Variation des Deutschen – Projekte und Perspektiven*. Berlin, Boston : de Gruyter, 519-540.
- Munro, R., S. Bethard, V. Kuperman, V. T. Lai, R. Melnick, C. Potts, T. Schnoebelen, H. Tily. (2010). Crowdsourcing and language studies : the new generation of linguistic data. *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. Association for Computational Linguistics, 122–130.
- Prokić, J., Ç. Çöltekin, J. Nerbonne (2012). Detecting Shibboleths. *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*. Avignon, France, 72-80.
- Sagot, B., K. Fort, G. Adda, J. Mariani, B. Lang. (2011). Un turc mécanique pour les ressources linguistiques : critique de la myriadisation du travail parcellisé. *TALN'2011*.
- Vaux, B., S. Golder. (2003). *The Harvard Dialect Survey*. Cambridge, MA : Harvard University Linguistics Department.
<http://www.snf.ch/en/researchinFocus/newsroom/Pages/news-150219-agora-funded-project-launch-of-voice-aapp.aspx>
<https://sites.google.com/site/adrianleemann/app-development>
<http://p3.snf.ch/project-164811>

¹ Joinson, Adam, ed. *Oxford Handbook of Internet Psychology*. Oxford University Press, 2007.

² Aragon, Yves, Sandrine Bertrand, Magali Cabanel & Hervé Le Grand. “Méthode d'enquêtes par internet : Leçons de quelques expériences”. *Décisions Marketing*, no. 19 (2000): 29-37. <http://www.jstor.org/stable/40592711>.

³ <https://crowdsourced.micropasts.org/>

⁴ <http://www.lesoir.be/826854/article/actualite/belgique/2015-03-19/quel-francais-belgique-parlez-vous>

⁵ <http://www.parlometre.ch/#!/about>

⁶ <http://www.donnezvotrefrancais.fr/>

⁷ <https://www.google.com/forms/about/>

⁸ <https://www.limesurvey.org/fr/>

⁹ <https://www.qualtrics.com/fr/>

¹⁰ Il s'agit ici de la version 2016-2017. Une nouvelle version, en cours de développement à la date de rédaction de cet article, est prévue pour un lancement mi-2018. Les modifications apportées à cette nouvelle version sont présentées en conclusion.

¹¹ <http://pybossa.com>

¹² Dans les enquêtes préalables, davantage d'informations géolinguistiques étaient demandées (déménagements, lieu d'habitation actuel, etc.), mais les cartes

linguistiques construites par la suite sont basées sur le lieu où le participant a passé la plus grande partie de son enfance (Avanzi et al. 2016), ce qui explique que ce soit le paramètre conservé ici pour permettre la comparaison.

¹³ Pour la période indiquée, 2879 participants ont répondu à l'ensemble des 12 questions, pour 3766 participants ayant répondu à la première question : nous avons donc un taux d'abandon au cours du quiz de 23,5%.

¹⁴ Ce sont notamment les commentaires que nous avons pu entendre lors de la fête de la Science, éditions 2016 et 2017 à Strasbourg, où le quiz linguistique était accessible au public, et où nous avons pu observer les réactions des participants (notons au passage que pour des soucis de comparabilité, les résultats ne tiennent pas compte des passations lors des fêtes de la Science). Voir Avanzi et Thibault (à par.) sur cette question des dénominations 70, 80 et 90 et des mythes qui leurs sont associés.

¹⁵ La Corse n'est pas représentée sur les cartes du quiz du fait que nous n'avons pas pu collecter un nombre suffisant de participants pour cette région. Pour chaque carte, il y a 5 niveaux de rouge, représentant 0-20%, 20-40%, 40-60%, 60-80%, 80-100% des participants (par département) respectivement, allant du blanc/beige au rouge foncé.

¹⁶ Voir récemment sur le *Midi Libre* (site web), mercredi 25 octobre 2017, l'article intitulé *Pain au chocolat ou chocolatine : qui remporte la bataille dans la région ?*, mais le débat n'est pas nouveau comme le montre cet autre article, daté de 2012 : *Vous êtes plutôt « pain au chocolat » ou « chocolatine » ?*, Sud Ouest (site web), mercredi 17 octobre 2012.

¹⁷ <http://wortschatz.uni-leipzig.de/en> (French mixed corpus based on material from 2012; Phrases: 74,823,426; Types: 7,873,935; Mots: 1,468,766,604).