Crowdsourcing Regional Variation Data and Automatic Geolocalisation of Speakers of European French

Jean-Philippe Goldman[1], Yves Scherrer[2], Julie Glikman[3], Mathieu Avanzi[4], Christophe Benzitoun[5], Philippe Boula de Mareüil[6]

[1] LATL-CUI University of Geneva, [2] University of Helsinki,
[3] LILPA University of Strasbourg, [4] FNRS/Université catholique de Louvain,
[5] ATILF Université de Lorraine, [6] LIMSI-CNRS

jeanphilippegoldman@gmail.com

Abstract

We present the crowdsourcing platform *Donnez Votre Français à la Science* (DFS, or "*Give your French to Science*"), which aims to collect linguistic data and document language use, with a special focus on regional variation in European French. The activities not only gather data that is useful for scientific studies, but they also provide feedback to the general public; this is important in order to reward participants, to encourage them to follow future surveys, and to foster interaction 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 geolocalisation system; i.e., a quiz that guesses the linguistic origin of the participant by comparing their answers with previously gathered linguistic data. For the geolocalisation activity, we set up a simulation framework to optimise predictions. Three classification algorithms are compared: the first one uses clustering and shibboleth detection, whereas the other two rely on feature elimination techniques with Support Vector Machines and Maximum Entropy models as underlying base classifiers. The best-performing system uses a selection of 17 questions and reaches a localisation accuracy of 66%, extending the prediction from the one-best area (one among 109 base areas) to its first-order and second-order neighbouring areas.

Keywords: language variation, regionalism, crowdsourcing, geolocalisation, linguistic geography, cartography

1. Introduction

Linguistic surveys and experiments can now easily reach thousands of people through the internet and smartphones. This 'crowdsourcing' methodology allows researchers to collect linguistic resources on a large scale (Cook et al. 2013), helping to describe linguistic phenomena (regional variation, non-normative forms, among others) that are underrepresented in traditional corpora. As a quick and easily reproducible way of collecting data, one can even conceive of using crowdsourcing over several years in order to compare successive snapshots of linguistic variables and thus describe variation in time.

In recent years, various crowdsourcing projects have been set up for collecting linguistic data through web applications, such as *Français de nos régions* (Avanzi et al. 2016) for European and Canadian French, *Atlas der deutschen Alltagssprache* (Möller & Elspaß 2015) for regional varieties of German, the *Harvard Dialect Survey* (Vaux & Bert 2013) for regional variation in the USA, *VerbaAlpina* (Krefeld and Lücke 2014) for dialects spoken in the Alps, and more recently, the twin websites *tonaccent/dindialäkt* focusing on the perception of Swiss French accents and Swiss German dialects (Goldman et al. 2018).

Other applications have been explicitly framed as geolocalisation games, aimed at predicting the provenance of the user on the basis of their naming of a set of objects. Whereas some projects have been distributed as smartphone apps, e.g. to document English and German dialectal variation (*Dialäkt Äpp*, Leemann et al. 2016; *English Dialects App*, Leemann et al. 2018a; *Grüezi, Moin, Servus*, Leemann et al. 2018b), other projects - mainly from outside the academic community - were designed as websites, e.g. the quizzes from the Belgian newspaper Le Soir (*«Quel français de Belgique parlez-vous?»*) or from the Télévision Suisse Romande (with *«le Parlomètre romand»*).

All these initiatives - based on such a popular theme as language variation - met with great success and collected large linguistic datasets, enabling novel scientific studies in this field.

In this paper, we present a crowdsourcing platform - *Donnez Votre Français à la Science* (DFS, or "*Give your French to Science*") - that includes several activities related to language variation. It is dedicated to French-speaking Europe (i.e., mainly Belgium, France and Switzerland[1]) and it is designed to be open to linguists to initiate their own projects. Moreover, it aims to deliver popularised feedback to the general public. After a general presentation of the platform, we present two activities: a linguistic quiz with immediate feedback to the participant, and a geolocalisation game - the first, to our knowledge, that covers French-speaking Europe. A large part of the paper is devoted to the development of the geolocalisation algorithm. The platform can be reached at this address: http://donnezvotrefrancais.fr

2. Crowdsourcing platform

The DFS platform is derived from PyBossa, a robust crowdsourcing framework that is designed for developing various interactive activities, based on the Flask micro-framework. The flexibility of the latter allowed for required adaptations such as having a limited number of questions/items per survey and a feedback mechanism, providing a score and/or a dynamically-computed map.

[1] Aosta Valley in Italy, despite having French as an official language, yielded too few participants.

Our ultimate goal is to help the linguists to set up their own survey, to collect data and to provide feedback to the participants. Our intention is that all collected data will eventually be made available on an open-source basis, making it accessible to third-party researchers.

Participants are invited to create an account for the platform (this can be facilitated via social networks) and provide basic sociolinguistic information (year of birth, gender, childhood location, and current location), following standard privacy cautions.[2] This information also allows us to entice contributors to return to the platform by informing them about new tasks. At the end of a task, the participant is invited to share their results on social media, thereby advertising the platform.

To this date, we have implemented a linguistic quiz (3.) and a geolocalisation task (4.) within the DFS platform. Both tasks give direct feedback to the participants. Such feedback is, in our view, very important, as it fosters the participant's understanding of the relevance of the task for research.

3. Linguistic quiz

In a similar vein to previous surveys on variation in European French (Avanzi et al. 2016), and on the basis of some of their data, we created a quiz in which the participants are asked to guess the meaning of regionalisms, in particular words or expressions. The quiz takes the form of a multiple choice questionnaire where a single answer is correct. An example concerning the regionalism *nareux* is given below:

(1) When one says about somebody that they are *nareux*, does this mean that that person:
   a. is picky with food?
   b. has a big nose?
   c. has a stuffy nose?
   d. has nausea?

The participant is immediately informed whether their selected answer is correct or wrong, and a short linguistic explanation is given, illustrated with a map. At the end of the quiz (i.e., after 12 questions), the participant is awarded a final score, which can be easily shared on social networks. This quiz has various aims. By its fun aspect (it is gamified with a final score, and the different answers proposed are often expressed in a humorous way), it is supposed to attract participants and make them aware of the platform in general. Through the explanations given during the quiz, it informs the participant about recent results of linguistic research. Finally, through the scores obtained by the participants, the linguist obtains a better picture of the vitality and passive knowledge of regionalisms in the population.

After a few weeks online, almost 3000 participants completed the quiz. 37% of these provided information about their linguistic origin (country and ZIP code). About the same proportion (39%) also gave sociolinguistic information such as birth year and gender. On average, participants correctly answered 6.7 out of 12 questions. Figure 1 depicts the distribution of scores for all participants.

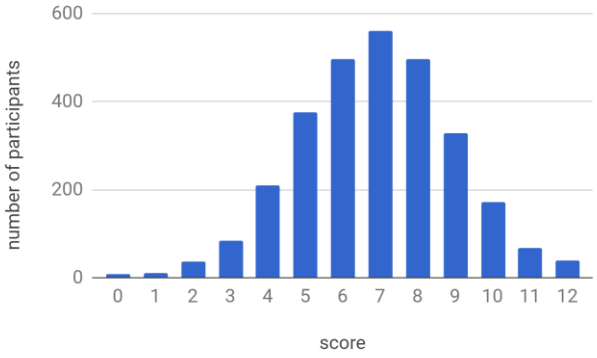

Figure 1. Distribution of scores (out of 12) for 2982 participants.

It is interesting to compare the quiz results, which represent passive knowledge about regionalisms, with results obtained in previous surveys (Avanzi et al. 2016), showing where the regionalisms are actively used (cf. the example questions in Section 4.1). Two patterns arise. Figure 2 demonstrates the first pattern, where the two maps are similar, meaning that a regionalism is not widely understood outside its area of active use.[3]

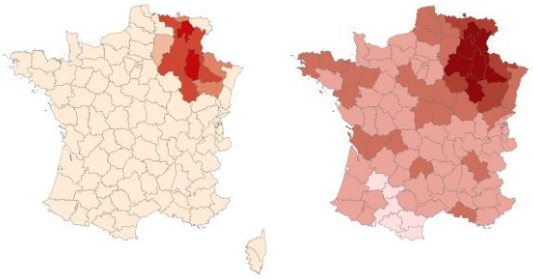

Figure 2. Active use (left, original survey with 8000 participants) vs. passive comprehension (right, crowdsourced quiz with 3000 participants) for the regionalism *nareux* 'picky'.

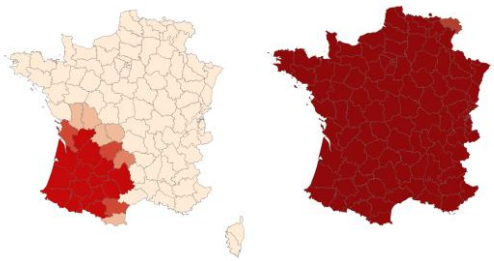

Figure 3. Active use (left, original survey with 12,000 participants) vs. passive comprehension (right, crowdsourced quiz with 3000 participants) for the regionalism *chocolatine* 'chocolate croissant'.

[2] The geolocalisation game can be played without creating an account or being signed in, in order not to raise suspicions that the sociolinguistic information will be used for the predictions.

[3] Corsica is not displayed on the quiz maps as we were not able to gather enough participants from this region.

Figure 3 illustrates the second pattern, where the two maps diverge, meaning that a regionalism is understood widely even though it is actively used only in a restricted area. In this particular case, the regionalism *chocolatine* had been the object of various lively discussions on social media over the last few years, resulting in nearly-universal comprehension in the whole French-speaking area of Europe.

Figure 4 plots the 12 quiz items according to their mean recognition rate (correct guesses) against the standard deviation of the guesses. While it shows a general tendency of lower variance with increasing mean scores (i.e., the better a regionalism is known by the participants, the higher the chances that it is known in the entire territory of inquiry), there are some noteworthy exceptions. The regionalism *péguer* 'to stick' has a fair recognition rate of about 0.6, but it has the highest standard deviation, meaning that its recognition rate varies widely across the area. On the contrary, the question about the number 80 yielded exceptionally low recognition scores, due to the particular way the question was asked.

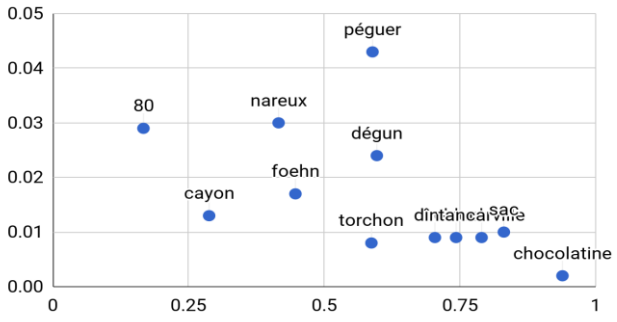

Figure 4. Mean recognition rate (horizontal axis) vs. standard deviation (vertical axis) for the 12 quiz items.

4. Geolocalisation

Using the same online surveys (Avanzi et al. 2016) as source material, we created another task focusing on geolocalisation. The main goal of this task was to provide a playful incentive to attract participants for further tasks on the DFS platform and to supplement the existing surveys with additional data points to continuously refine the accuracy of the geolocalisation task. At the same time, the semi-automatic selection of items for the task provided an interesting application of various methods of data analysis and machine learning.

There have been two major approaches to speaker geolocalisation (or dialect identification) in the literature: the **corpus-based approach** predicts the dialect of any text fragment extracted from a corpus; this approach has been followed by the VarDial shared tasks in recent years (e.g. Malmasi et al. 2016, Zampieri et al. 2017), but also by Scherrer & Rambow (2010) or Rahimi et al. (2017), for example. The **dialectological approach** tries to identify a small set of distinguishing dialectal features, which are then elicited interactively from the user in order to identify their dialect (Leemann et al. 2016, 2018a, 2018b). The task proposed here follows the dialectological approach.

The problem of geolocalisation consists in predicting the dialect/regiolect of a speaker (typically, a speaker that has not participated in the survey) by asking a set of questions (typically about a small subset of the surveyed variables). Given our motivations, the success of a geolocalisation method should not only be assessed in terms of the percentage of correct predictions, but also by its ability to entertain and surprise participants. Three parameters influence this success:

N - the number and type of questions to be asked. No more than 20 questions should be asked to keep the required attention span short.

M - the number of the areas to predict. The areas should reflect the relative scarcity of regional variation in current French, but too-large areas could make the problem look trivial and uninteresting.

A - the accuracy of the predictions. The method should obviously make as good predictions as possible, but we estimate that about 2/3 of correct predictions is required for a sustainable level of participant involvement.

In the following sections, we give some details about how we approached this optimisation problem. In other words, we wish to select the best set of questions from the previously collected survey data (with *N* being as low as possible) with the best set of prediction areas (i.e., the *M* areas being as small as possible), in order to achieve the highest accuracy *A*. In order to estimate the success of a crowdsourced geolocalisation task before its launch, we set up a simulation framework to find the optimal parameter settings.

4.1 Data

We rely on data from two surveys on regionalisms in European French (France, Belgium and Switzerland), which were carried out in 2015-2016 as a part of the project *Français de nos régions* (Avanzi et al. 2016); key information regarding the surveys is shown in Table 1.

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

Table 1: Number of questions and participants in the two surveys.

Each participant was asked 40 or 90 multiple-choice questions on lexical regionalisms (small parts of the surveys also concerned morpho-syntactic and phonological variation). Some questions were illustrated by pictures. They could be direct questions of word usage (see question (2) below) or they might encompass a definition of a concept or an object (see question (3)). The number of possible answers varied from 2 to 11, and multiple answers were allowed.

(2) Do you use the word *s'entrucher*?
  a. yes
  b. no

(3) How do you call the piece of cloth that is used to wash the floor?
   a. serpillière
   b. torchon
   c. since
   d. wassingue
   e. loque
   f. pièce
   g. panosse
   h. toile
   i. chiffon
   j. lave-pont
   k. patte

4.2 Simulation framework

We applied two important pre-processing steps to the survey data. First, we settled on a set of 109 administrative areas as an upper bound for *M*: we considered 96 French departments, 7 Swiss cantons (of the French-speaking part of Switzerland, called *Romandie*), and 6 Belgian provinces (of the French-speaking part of Belgium, called *Wallonie-Bruxelles*). Although survey participants provided ZIP code information, we aggregated the subjects into 109 areas to avoid data scarcity issues in sparsely populated areas. Second, we matched participants from Survey 1 with participants from Survey 2 on the basis of their origin, leading to a total dataset of 6463 participants.

In order to evaluate different settings of the parameters *N*, *M* and *A*, we set up a simulation framework using solely the survey data in a leave-one-out fashion. The general idea is to train a model on the aggregated data of all except one participant, predict the origin of the left-out participant, and compare the prediction with the ground truth. However, contrary to a true leave-one-out setting, we chose not to remove the test participant from the training data for efficiency purposes (avoiding the need to train a new model for each participant). As the training data was aggregated and contains more than one participant for each area and question, there was never exactly the same data point in the training and test corpus, allowing us to take this methodological shortcut.

We considered two approaches to find the best parameter settings for geolocalisation, one based on clustering and shibboleth detection, and one based on feature elimination.

4.3 Clustering and shibboleth detection

This approach consisted of two steps: we first determined an optimal areal partition using hierarchical clustering, and then applied the shibboleth detection algorithm of Prokić et al. (2012) to find the most characteristic set of questions for each area.

Figure 5 shows an example of hierarchical clustering solutions using Ward's method and 10 target clusters, obtained using the complete dataset (6463 participants, 130 questions). It is worth pointing out that the aggregated data clusters nicely into geographically coherent and linguistically sensible regions, suggesting that the quality of the survey data is good.

Figure 5. Resulting areas after applying hierarchical clustering with Ward's method and ten target clusters.

The shibboleth detection algorithm was then used to list the five most characteristic linguistic variants ("shibboleths") per cluster. For example, it produced the variants *encoubler, septante, nonante, ça joue, souper* for the French Swiss area (light green area of the map), or *péguer, challer, soixante-dix, sèche-cheveux, quatre-vingt-dix* for the Provence area (cyan area of the map).

The success of this approach was limited in our case, owing essentially to two factors. First, it was very sensitive to the clustering parameters: a slight change in the number of target clusters, or a change in the clustering algorithm, led to considerable differences in simulation performance. As there are no universally applicable criteria for evaluating the quality of a hierarchical clustering, this essentially amounted to a trial-and-error process with little scientific value. Second, the core assumption of the shibboleth detection algorithm, namely that there are linguistic variants whose geographic distribution coincides with a cluster, did not seem to hold in our data. While clear regionalisms exist for Switzerland and Belgium (and are successfully identified, as shown above for the Swiss case), the inferred clusters within France are much less clearly correlated with single linguistic variants, but rather emerge through the combination of a large number of gradual linguistic differences. Due to these problems, we did not pursue this approach any further.

4.4 Feature elimination

In a second approach, we did not fix a geographical partition in advance, but kept the 109 areas as defined above while finding the optimal set of questions. For this, we applied feature elimination techniques, as detailed below. Once the questions were determined, we dynamically expanded the predictions to n-best areas or neighbours. The approach is summarised in the four following steps:

1. As the linguistic variables could have several variants with different distributions, we treated each variant separately and binarised the data from 130 n-ary variables to 639 binary variables. For example,

question (3) of section 4.1 represents an 11-ary variable (11 possible answers). This variable was converted into 11 binary variables of the form "Do you call the piece of cloth *serpillière*?", "Do you call the piece of cloth *loque*?", etc.

2. Some variants were hardly ever used or showed no geographic variation at all, so we discarded them with a single-pass feature elimination based on χ^2 score. In other words, we removed those variables that were the least statistically dependent on location. We found a lowest average distance between prediction and ground truth with 150 variants as shown in Figure 6, and settled on this value for the following steps. For the floor-cloth example, this step eliminated the variants *chiffon*, *patte* and *lave-pont*, which very few survey participants had selected.

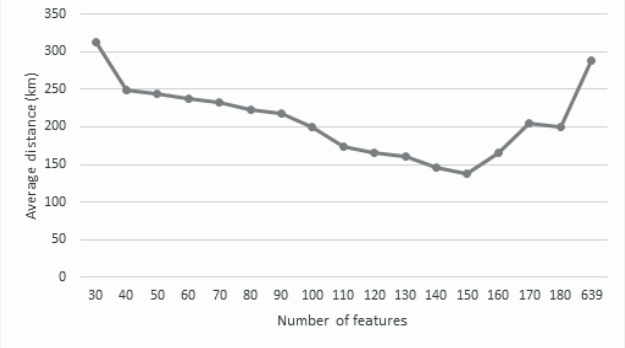

Figure 6. Average distance (in km) between predicted and actual areas as a function of the number of variables.

3. While we could have continued using the χ^2 method of step 2 to further reduce the number of variables to an acceptable value (recall that we aim for a value of N close to 20), we opted for recursive feature elimination techniques (RFE) instead, in order to take into account the dependencies between variables. Therefore, starting with the 150 binary variables of step 2, we trained a classifier and used RFE (Guyon et al. 2002) to repeatedly remove the variant that contributed least to the classification. We ran parallel experiments with two classification algorithms, SVM and MaxEnt. Both classifiers achieved much better simulation results than the χ^2 method, with SVM performing slightly better than MaxEnt (see Figure 7). We found that the χ^2 feature elimination - because it looks at each variable independently - ended up proposing a lot of variables that predict the same regional partition (e.g. Switzerland vs. France+Belgium, which is the most salient one), whereas the RFE methods yielded more complementary sets of variables. For the following steps, we settled on a smaller window of 10-40 binary variables (instead of 0-150 after step 2), as obtained by the SVM or MaxEnt methods.

4. We evaluated the simulation results in several ways. Figures 6 and 7 show average distances between the centroids of the predicted and true areas. A simpler measure is area accuracy (i.e. whether the true area has been correctly predicted or not), which is also reported below (Figure 8). We also extended area accuracy to immediate neighbours (i.e. whether the true area is equal to the predicted area or one of its neighbouring areas) and second-order neighbours.

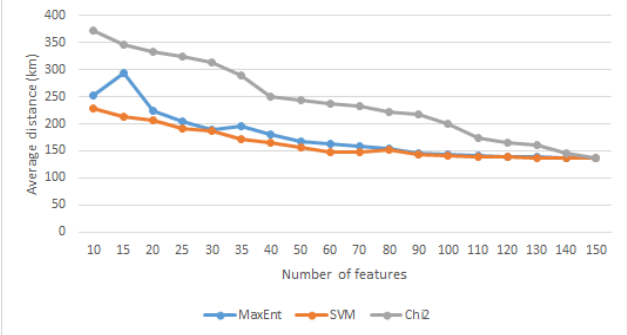

Figure 7. Average distance (in km) between predicted and actual areas as a function of the number of retained variables for SVM and MaxEnt classifier RFE and χ^2 feature elimination.

Figure 8 shows simulation results for both classifiers from which we can draw a few conclusions: first, the results stabilised at around 40 binary variables, i.e. about 30 n-ary questions. Second, extending the predictions to first-order neighbours improved accuracy by +30%, while extending them to second-order neighbours added a further +20%. With 20 variants (representing 17 n-ary questions), the accuracy score was 66.2% on second-order neighbours. This setting satisfies our target values for the variables *A* and *N*.

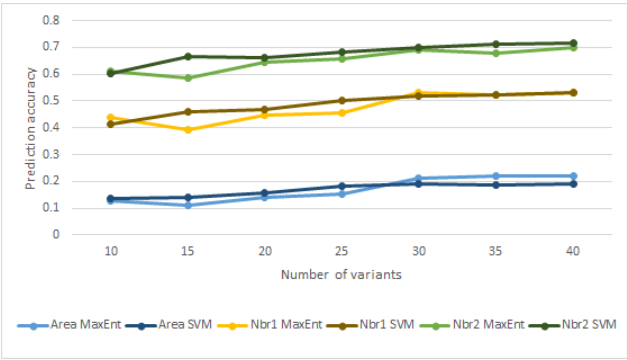

Figure 8. Prediction accuracy from 10 to 40 variants with SVM and MaxEnt classifiers and considering exact area accuracy, immediate neighbouring areas (Nbr1) and second-order neighbours (Nbr2).

4.5 Crowdsourcing implementation

The geolocalisation quiz was implemented in the DFS platform using the 15 most relevant n-ary questions as obtained by the MaxEnt and SVM RFE approaches.

With a sharing mechanism on social networks and media coverage, we were able to gather data from 8000 participants, who were led alternatively to the MaxEnt or SVM surveys. Later, we added a third survey based on a manual variant selection of 15 questions, which 500 participants completed.

Whatever the version of the quiz (MaxEnt, SVM or manual selection), a probability was computed for each of

the 109 areas after the 15 questions, and was displayed on a result map of the European French-speaking area, as in Figure 9.

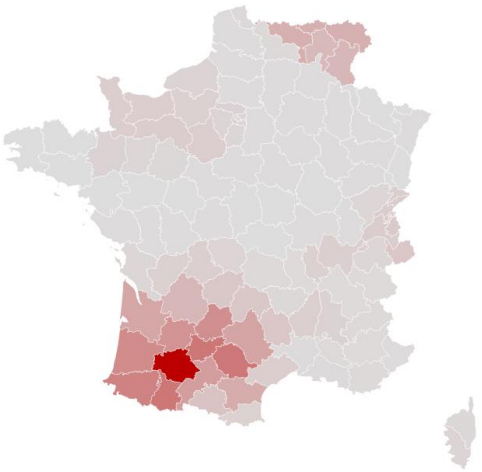

Figure 9. Example of result map of the geolocalisation presented to the participant.

Participants were also then asked for sociolinguistic information (country and ZIP code, age, gender). About 40% of participants provided these data.

4.6 Comparative results

Table 2 compares the classification results from the simulations (as in 4.4) with those for the real participants (as explained in section 4.5) who gave their true localisation information. With both automatic methods, we reached the desired accuracy threshold with comparable area sizes and number of variables (about 20). However, the variables selected by the SVM classifier intuitively corresponded better to the variation patterns observed in the original survey data.

	Part	Best	5-Best	Nbr-1	Nbr-2
RFE MaxEnt	simulated	14 %	49 %	47 %	64 %
	crowdsourced	11 %	43 %	40 %	62 %
RFE SVM	simulated	13 %	46 %	46 %	66 %
	crowdsourced	13 %	47 %	47 %	64 %
Manual selection	simulated	10 %	36 %	40 %	57 %
	crowdsourced	5 %	16 %	12 %	18 %
Random		<1%	4.5 %	4.5%	9%

Table 2. Geolocalisation results with crowdsourced and simulated data (percentages are f-scores for 109 areas).

4.7 Discussion

Our attempt to apply machine learning techniques for question (and area) selection led to a 66% correct-response rate (as in Table 2 with RFE SVM extended to second-order neighbours) which was confirmed with crowdsourced real data (64% of correct responses). The two main advantages of this automatic approach consist in optimising the selection of questions and estimating the success of a crowdsourced linguistic campaign before launch.

Although originally intended as an optimisation method for defining and optimising large areas in a region-guessing activity, we also ended up with a more fined-grained localisation (areas instead of clusters of areas) with a colour-scaled probability.

One major drawback of this approach is the dependency on earlier surveys for variable selection and simulation. Also, it proved difficult to convince the participants to fill out their sociolinguistic information after displaying the result of the quiz. As mentioned above, we did not want to prompt them to fill out the questionnaire beforehand in order not to raise suspicions.

5. Conclusion

This paper presents a new web platform for hosting activities such as linguistic surveys, including different types of foreseeable questions (text, picture or sound) and different types of answers (multiple-choice, free text, or sound recording). Although generic survey platforms already exist on the web, they are not well adapted for linguistic surveys.

We also present two activities that have been implemented on this web platform: a linguistic quiz about regionalisms and a geolocalisation task. For the latter, we compared several approaches for defining an optimised set of questions and areas. In a simulated setup, we found that a recursive feature elimination approach using a MaxEnt base classifier worked best, whereas the result of the crowdsourcing campaign showed a slight advantage for the SVM base classifier. In both cases, predicting a single area out of 109 proved difficult, but accuracy levels approached 66% when including both first- and second-order neighbours in the prediction. Also, automatic approaches to question selection turned out to work better than a linguistically informed manual selection, although the crowdsourced results for the latter should be taken with a pinch of salt due to the relatively low number of participants.

The presented framework could easily be localised and adapted to other languages. Also, provided that source surveys are available, it can easily be adapted to other French-speaking areas, such as parts of Africa or Canada, with minor adaptations to the maps.

Our next task is to extend the platform so that linguists can set up their own surveys. Moreover, additional information will be collected from the participants, such as educational level and job typology, in order to compare diatopic and diastratic variation patterns.

6. Acknowledgements

This research is funded by DGLFLF (Délégation générale à la langue française et aux langues de France) and supported by these academic partners: University of Geneva (Switzerland), University of Helsinki (Finland), UCLouvain (Belgium), University of Strasbourg (France), LIMSI-CNRS (France), and ATILF-Université de Lorraine (France).

7. Bibliographical References

Avanzi, M., C. Barbet, J. Glikman, J. Peuvergne (2016). *Présentation d'une enquête pour l'étude les régionalismes du français.* Actes du 5ème congrès mondial de linguistique française (CMLF). Tours, France, 1-15.

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.

Goldman J.Ph. et al. (2018). *Strategies and Challenges for Crowd-Sourcing Dialect Perception Data [for Swiss German and Swiss French]*. LREC Conference, Miyasaki, Japan.

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). *VerbaAlpina - Der alpine Kulturraum im Spiegel seiner Mehrsprachigkeit.* Ladinia XXXVIII, 189-211.

Leemann, A., M.-J. Kolly, R. Purves, D. Britain, E. Glaser (2016). *Crowdsourcing language change with smartphone applications.* PLOS ONE.

Leemann, A., M-J. Kolly, D. Britain (2018a). *The English Dialects App: the creation of a crowdsourced dialect corpus.* Ampersand 5, 1-17.

Leemann, A., S. Elspaß, R. Möller, T. Grossenbacher (2018b). *Grüezi, Moin, Servus – Wie wir wo sprechen.* Hamburg: Rowohlt.

Malmasi, S., M. Zampieri, N. Ljubešić, P. Nakov, A. Ali, J. Tiedemann (2016). *Discriminating between Similar Languages and Arabic Dialect Identification: A Report on the Third DSL Shared Task.* Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3), Coling. Osaka, Japan, 1-14.

Möller, R., S. Elspaß (2015): *Atlas zur deutschen Alltagssprache*. In: Kehrein, R., A. Lameli, S. Rabanus (eds.): *Regionale Variation des Deutschen – Projekte und Perspektiven.* Berlin, Boston: de Gruyter, 519-540.

Prokić, J., Ç. Çöltekin, J. Nerbonne (2012). Detecting Shibboleths. Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH. Avignon, France, 72-80.

Rahimi, A., T. Baldwin, T. Cohn (2017). *Continuous Representation of Location for Geolocation and Lexical Dialectology using Mixture Density Networks.* Proceedings of EMNLP. Copenhagen, Denmark, 167-176.

Scherrer, Y., O. Rambow (2010). *Word-based Dialect Identification with Georeferenced Rules*. Proceedings of EMNLP. Cambridge, MA, 1151-1161.

Vaux, B., S. Golder (2003). *The Harvard Dialect Survey*. Cambridge, MA: Harvard University Linguistics Department.

Zampieri, M., S. Malmasi, N. Ljubešić, P. Nakov, A. Ali, J. Tiedemann, Y. Scherrer, N. Aepli (2017). Findings of the VarDial Evaluation Campaign 2017. Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial), EACL. Valencia, Spain, 1-15.