CROWDSOURCED MAPPING OF PRONUNCIATION VARIANTS IN EUROPEAN FRENCH

Yves Scherrer*, Philippe Boula de Mareüil§, Jean-Philippe Goldman*

* LATL-CUI, University of Geneva, Switzerland; § LIMSI-CNRS, Orsay, France
yves.scherrer@unige.ch; philippe.boula.de.mareuil@limsi.fr; jean-philippe.goldman@unige.ch

ABSTRACT

This study aims at renewing traditional dialectological atlases to provide a mapping of pronunciation variants by using crowdsourcing. Based on French spoken in France, Belgium and Switzerland, it focuses on mid vowels whose quality may be open or close and shibboleths such as final consonants which may be maintained or deleted, as a function of speakers' background. Over 1000 subjects completed a questionnaire in which they were asked which one of the two possible pronunciations was closer to their most usual pronunciation. Their responses for 70 French words are displayed in the form of maps. This graphical layout enables the general public and phoneticians to readily visualise where phonological constraints such as the "loi de position" are violated.

Keywords: linguistic atlases, diatopic variation, French pronunciation

1. INTRODUCTION: MOTIVATION

The study of regional variation in pronunciation patterns is one of the oldest areas of phonetics and dialectology. In particular, large-scale linguistic surveys were carried out for many languages during the early 20th century. Most of these surveys contain extensive sections on phonetics and phonology. However, the resulting atlases only partially reflect today's speech.

A case in point is European French. The *Atlas linguistique de la France* (ALF) [11] maps language varieties spoken in the late 19th century. In the second half of the 20th century, a series of follow-up projects were undertaken to document dying dialects in more detail; these projects are subsumed under the title *Atlas linguistique de la France par régions* (ALFR) [23] (see also [12]). Today, these dialects have mostly been replaced by regional French varieties. Unfortunately, no large-scale survey of current regional French has been endeavoured yet.

There have been proposals to fill this gap, since Walter's [26] phonological investigation. Within the framework of the *Phonologie du français contemporain* (PFC) project [5, 6, 7], samples of spoken French covering a number of Francophone survey points (today 36) have been collected. Even though the data were used in various studies on regional variation [3, 4, 22], the sparse inquiry point network does not allow us to create an atlas in the form of the ALF. The TANLAF proposal [14] is more ALF-like, but its coverage is currently limited to urban areas of the northern third of France.

The goal of the present study is to examine whether online surveys, or more generally crowdsourcing [9], can be used so as to collect linguistic data in sufficient quality and regional diversity, with a fine-grained geographic coverage. Based on European French (i.e., the French language spoken in France, Belgium and Switzerland), this work focuses on the pronunciation of words which typically exhibit diatopic (geographic) variation. It is inspired by two surveys which have been conducted successfully in the last few years: the *Harvard Dialect Survey* (HDS) for American English [24], and the *Atlas der deutschen Alltagssprache* (AdA) for German [8]. These atlases, which involved thousands of participants, primarily address lexical variation but they include phonetic variation issues as well.

The next section introduces the sources of phonetic variation we investigated for European French. Section 3 presents the questionnaire we developed for collecting the data. Section 4 provides some results in the form of maps, Section 5 discusses the crowdsourcing-based methodology and considers future work.

2. PHONETIC VARIATION IN EUROPEAN FRENCH

Several sources of diatopic phonetic variation at the word level have been identified: mid vowels, nasal vowels, the schwa vowel, the front/back /A/, vowel quantity, glides, final consonants… Given the merger of phonological oppositions related to the back /ɑ/ and the nasal vowel /œ̃/ in standard (Parisian) French [10, 15, 19, 18], we found it too difficult to get reliable information from the general public regarding these vowels. We also

found vowel quantity, the schwa realisation/deletion and the glide diaeresis/synaeresis too speech rate-dependent. Hence, we focused on three pairs of mid vowels /e/~/ɛ/, /ø/~/œ/, and /o/~/ɔ/, but also included shibboleths which are emblematic of Belgian French, in particular the use of [w] (instead of the canonical /v/ or /ɥ/ in some words) and the realisation of final consonants.

The "loi de position", which accounts for the tendency of mid vowels to be open in closed syllables and close in open syllables, does not apply in a straightforward manner to French [2, 25]. This constraint is often claimed to better apply to southern French, where the /ɔ/ tends to be close in words such as *botté* 'booted' and open in words such as *sauf* 'except' [3, 4]. Instead, some eastern French speakers may distinguish words such as *peau~pot* (/o/~/ɔ/) 'skin'~'pot' and pronounce *jeune* 'young' with a close /ø/. This led us to build up a list of 70 words whose pronunciations may vary as a function of the speaker's region. Further details are given in the following section.

3. THE QUESTIONNAIRE

A web-based questionnaire was designed using the Labguistic platform [21]. It is made up of three parts: in the first part, information about the participant is collected; the second part contains the actual phonetic questions; in the third part, the participant may provide feedback about the experiment. Completing the entire questionnaire takes about 10-15 minutes.

In the first part, the participant is asked about his/her linguistic curriculum:

- in which city the participant currently lives;
- in which department (France), province (Belgium) or canton (Switzerland) the participant spent most of his/her life;
- in which department, province or canton the participant grew up.

This information allows us to pool participants on three levels of granularity: on the city level (for the current residence only), on a fine-grained areal level (department/province/canton) and on a coarse-grained areal level (the 22 official regions for France, one region for the French-speaking part of Belgium and one region for the French-speaking part of Switzerland). Additional information is recorded about the participant, such as age and gender, and whether (s)he is a native speaker of French.

The second part of the questionnaire consists of the 70 items. For each item, displayed in French orthography, the participant hears a recording with two possible pronunciations (e.g., for the word *chose* 'thing', one with an open vowel and one with a close vowel). The participant is instructed to choose whether the first or the second pronunciation is closer to his/her own most usual pronunciation, by clicking on the button marked **1** or **2**. A third button may be clicked by the participant if (s)he is unable to perceive a difference between the two pronunciations (in case of phonological deafness). The items as well as the two pronunciations per item are presented in random order. The participants may listen to the recordings as many times as they want. The samples were recorded by a phonetician from Paris, who did not utter final schwas.

According to the brief description given in the previous section, the 70 items cover the following phonetic contexts:

- /e/~/ɛ/ in open syllable (e.g. *parfait* 'perfect'): 19 words, 6 of which form minimal pairs in standard French (e.g. *épée~épais* 'sword'~'thick');
- /e/~/ɛ/ in closed syllable (e.g. *père* 'father'): 4 words;
- /o/~/ɔ/ in open syllable (e.g. *sot* 'foolish'): 2 words;
- /o/~/ɔ/ in closed syllable (e.g. *chose* 'thing'): 28 words, 2 of which form a minimal pair in standard French;
- /o/~/œ/ in open syllable: 1 word (*social* 'social') to exemplify 'o'-fronting [1, 3, 17, 20];
- /ø/~/œ/ in closed syllable (e.g. *chanteuse* 'singer'): 10 words, 2 of which form a minimal pair in standard French (e.g. *jeûne~jeune* 'fasting'~'young');
- initial /w/: 2 words (*wagon* 'wagon', *huit* 'eight');
- final /t/: 2 words (*soit* 'either' , *vingt* 'twenty');
- final /s/: 2 words (*moins* 'less', *encens* 'incense').

In the final part of the questionnaire, the subjects are asked about the difficulty of the task: for instance, if in some items both pronunciations sounded unfamiliar to them.

4. THE PARTICIPANTS

The survey was launched on October 1, 2014. It was advertised through mailing lists, social networks as well as personal and professional contacts. Also, university teachers in the relevant domains were contacted to pass the information on to their students.

Within a few months, 1250 participants completed the survey and gave exploitable localisation information. 72% of participants are females and 42% work in language sciences. The age distribution of infor-

mants is the following: 12% under 20, 40% in their twenties, 15% in their thirties, 11% in their forties, 12% in their fifties and 10% over 60.

The number of participants currently living in rural areas turned out to be very low, which had two consequences on our analyses. First, rather than focusing on the place of residence, we focused on the place where the participant spent most of his/her life. Second, we focused on the coarse-grained areal level to avoid data sparsity. Figure 1 shows the distribution of participants according to these criteria.

We did not explicitly restrict our questionnaire to Europe. In particular, we collected responses of 28 speakers of Canadian French and 6 speakers from the French Overseas departments. We do not consider these responses here since the questionnaire was not specifically suitable for these varieties of French. We also discarded non-native speakers.

According to their feedback at the end of the questionnaire, 94% of participants found it easy to decide on all or nearly all items. This figure is confirmed by the percentage of clicks on the third button, which was used for 1.74% of answers. These figures suggest that people are able to perceptually distinguish between two pronunciations and that they are able to associate one of them with their own pronunciation.

5. RESULTS

In the following, we discuss the results obtained for the tested phonetic variables and illustrate them with frequency maps. Each map relates to one word and shows, for each region, what percentage of participants preferred the close vowel pronunciation. All maps use the same scale consisting of 5 equal relative frequency intervals. The maps were generated using ArcGIS. Some results are described below:

- /e/~/ɛ/ in open syllable. For most items, close vowels show up in the Southwest and in the northmost region of France, compared to more open vowels in the rest of the area. See Figure 2 for an example.
- /e/~/ɛ/ in closed syllable. There is almost no regional variation for these items; vowels are always open.
- /o/~/ɔ/ in open syllable. This vowel is predominantly close in France, whereas it is open in Belgium and in the northern parts of French-speaking Switzerland. The adjacent departments in Franche-Comté also show open vowels. See Figure 3.
- /o/~/ɔ/ in closed syllable. Here, we obtain open vowels in the South of France and in the northmost region of France. See Figure 4. For the items *grosse* 'fat' and *fosse* 'pit', Belgium aligns with the South, even more markedly than the North of France.
- /ø/~/œ/ in closed syllable. These items form two distinct classes: the first class, containing words such as *neutre* 'neutral' or *chanteuse* 'singer' follow the standard /ø/ pattern; the second class, containing *aveugle* 'blind' and *gueule* 'mouth', is pronounced with an open /œ/ throughout France and Belgium, but less so in French-speaking Switzerland (Figure 5). Within this class, there are considerable lexical and age-related effects, with older partic-

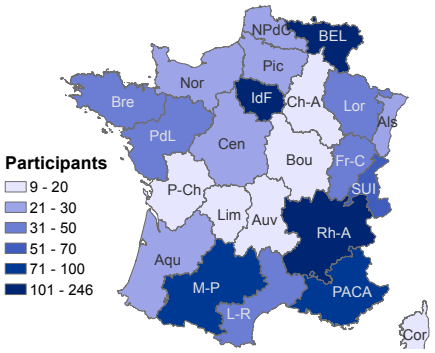

Figure 1: Numbers of participants per region in which they spent most of their lives.

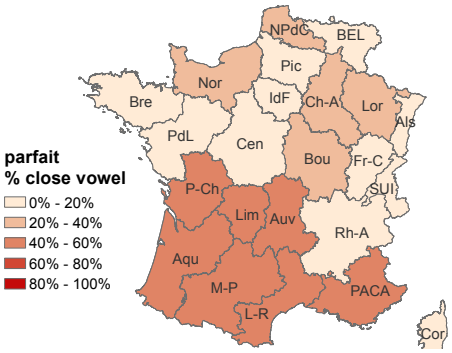

Figure 2: Distribution of /e/~/ɛ/ (darker~lighter) in *parfait* 'perfect'.

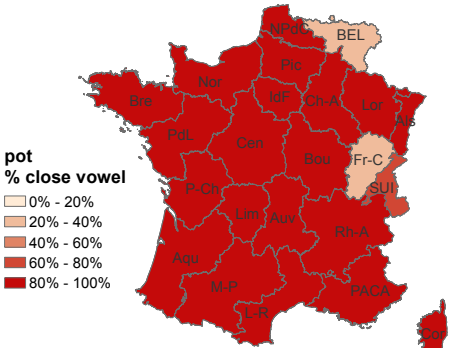

Figure 3: Distribution of /o/~/ɔ/ (darker~lighter) in *pot* 'pot'.

ipants using close pronunciations more often than younger participants.
- /o/~/œ/. The resulting map is rather uniform, showing a preference for the /o/ variant, without clear regional differences;
- For *wagon* 'wagon', *huit* 'eight' and *soit* 'either', Belgium departs from the other countries, as expected. The situation is less clear-cut for *vingt* 'twenty', where the final /t/ is pronounced in the whole Northeast, ranging from Belgium over Lorraine to Switzerland (see Figure 6).
- The word *moins* 'less' is pronounced with a final [s] in the Southwest, while *encens* 'incense' keeps its final [s] also in some cantons of Switzerland.

6. CONCLUSION AND FUTURE WORK

Crowdsourcing, rather quickly, allowed us to visualise phonetic variation across 70 words in French spoken in Europe. It proved possible to obtain about 1000 completed questionnaires in only two months, and more than 1200 questionnaires in half a year. This paradigm can thus be considered an effective, low-cost method to collect linguistic data and map diatopic variation. The maps shown in this article present a first glimpse on the collected data, but the latter deserve more in-depth analyses: we would like to investigate the effects of participants' age, education and mobility. Furthermore, we may aggregate data from various items and features to provide a more generic picture of the phonetic processes involved and perform dialectometrical analyses (e.g., along the lines of Goebl [12]). The collected material will also allow us to automatically predict the localisation of new speakers.

Compared with other crowdsourcing tasks, surveying regional language variation requires the participants to be evenly spread out over the investigated territory, making it challenging to get sufficient numbers of informants for some areas. As mentioned earlier, few responses were obtained in less-populated French departments, especially in the Limousin and Auvergne regions. Unfortunately, these regions precisely divide the North and the South of France. Hence, we are currently trying to recruit more participants around this major isogloss.

In addition, we plan to set up a follow-up questionnaire which could extend the current one in the following directions:
- Include non-European French varieties. This means that, for many items, there would be more than two possible answers, which would require to change the questionnaire setup (and, probably, the wordlist).
- Include non-phonemic variation patterns: there is known prosodic and lexical variation, which could be surveyed in the same fashion.
- Compare what speakers declare concerning their pronunciations and what they actually produce (or what other speakers produce, in the PFC database, for example). We could specifically add voice recordings to online questionnaires, along the lines of [13] or [16].

The same approach can apply to other languages.

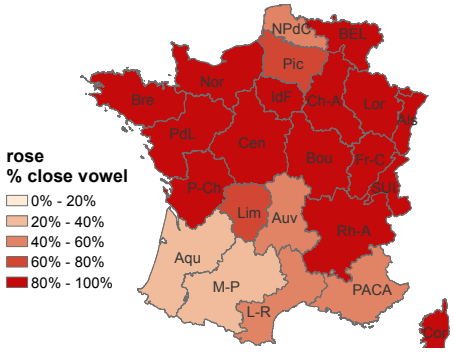

Figure 4: Distribution of /o/~/ɔ/ (darker~lighter) in *rose* 'rose'.

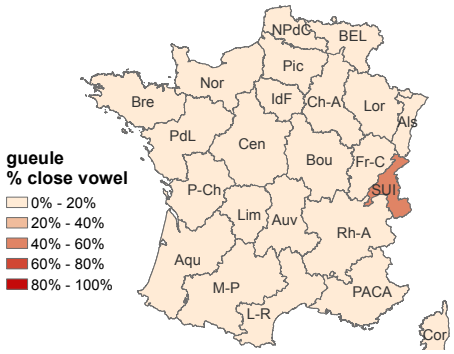

Figure 5: Distribution of /ø/~/œ/ (darker~lighter) in *gueule* 'mouth'.

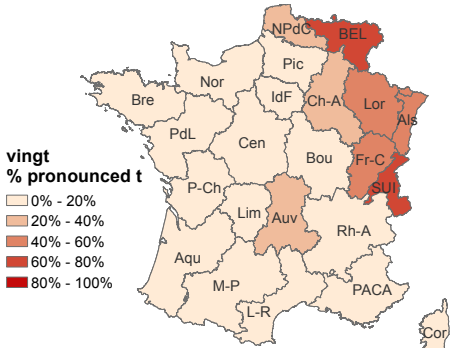

Figure 6: Distribution of final /t/ (pronounced~silent: darker~lighter) in *vingt* 'twenty'.

7. REFERENCES

[1] Armstrong, N., Low, J. 2008. C'est encœur plus jeuli, le Mareuc : some evidence for the spread of /ɔ/-fronting in French. *Transactions of the Philological Society* 106(3), 432–455.

[2] Armstrong, N., Pooley, T. 2010. *Social and linguistic change in European French*. Basingstoke: Palgrave Macmillan.

[3] Boula de Mareüil, P., Woehrling, C., Adda-Decker, M. 2013. Contribution of automatic speech processing to the study of Northern/Southern French. *Language Sciences* 39, 75–82.

[4] Dufour, S., Nguyen, N., Frauenfelder, U. H. 2007. The perception of phonemic contrasts in a non-native dialect. *Journal of the Acoustical Society of America Express Letters* 121, 131–136.

[5] Durand, J., Kristoffersen, G., Laks, B. 2014. *La phonologie du français : normes, périphéries, modélisation*. Nanterre La Défense: Presses Universitaires de Paris Ouest.

[6] Durand, J., Laks, B., Lyche, C. 2002. La phonologie du français contemporain : usages, variétés et structure. In: Pusch, C. D., Raible, W., (eds), *Romance corpus linguistics – Corpora and spoken language*. Tübingen: Narr 93–106.

[7] Durand, J., Laks, B., Lyche, C., (eds) 2009. *Phonologie, variation et accents du français*. Paris: Hermès.

[8] Elspaß, S. 2007. Variation and Change in Colloquial (Standard) German – The *Atlas zur deutschen Alltagssprache* (AdA) Project. In: Fandrych, C., Salverda, R., (eds), *Standard, Variation und Sprachwandel in germanischen Sprachen/Standard, Variation and Language Change in Germanic Languages*. Tübingen: Narr 201–216.

[9] Eskenazi, M., Levow, G.-A., Meng, H., Parent, G., Suendermann, D., (eds) 2013. *Crowdsourcing for Speech Processing: Applications to Data Collection, Transcription and Assessment*. Chichester: Wiley.

[10] Fónagy, I. 2003. Le français change de visage ? *Revue Romane* 24(2), 225–254.

[11] Gilliéron, J., Edmont, E. 1902–1910. *Atlas linguistique de la France*. Paris: Champion.

[12] Goebl, H. 2002. Analyse dialectométrique des structures de profondeur de l'ALF. *Revue de linguistique romane* 66(261-262), 5–63.

[13] Goldman, J.-P., Leemann, A., Kolly, M.-J., Hove, I., Almajai, I., Dellwo, V., Moran, S. 2014. A crowdsourcing smartphone application for Swiss German: Putting language documentation in the hands of the users. *Ninth International Conference on Language Resources and Evaluation*. Reykjavik. 3444–3447.

[14] Hall, D. J. 2012. Vers un Nouvel Atlas Linguistique de la France: aspects de méthodologie sociolinguistique et dialectologique. *Third International Congress on French Linguistics*. Lyon. 2171–2189.

[15] Hansen, A. 2001. Lexical diffusion as a factor of phonetic change: The case of Modern French nasal vowels. *Language Variation and Change* 13, 209–252.

[16] Kolly, M.-J., Leemann, A. to appear. Dialäkt Äpp: communicating dialectology to the public – crowdsourcing dialects from the public. In: Leemann, A., Kolly, M.-J., Dellwo, V., Schmid, S., (eds), *Trends in Phonetics and Phonology: Studies from German-speaking Europe*. Bern: Peter Lang.

[17] Landick, M. 1995. The mid-vowels in figures: Hard facts. *The French Review* 69(1), 88–102.

[18] Malmberg, B. 1966. *La phonétique*. Paris: Presses Universitaires de France.

[19] Malécot, A., Lindsay, P. 1976. The neutralization of /ɛ̃/~/œ̃/ in French. *Phonetica* 33, 45–61.

[20] Martinet, A. 1958. C'est jeuli, le Mareuc ! *Romance Philology* 11, 345–355.

[21] Ménétrey, P., Schwab, S. to appear. Labguistic: a web platform to design and run speech perception experiments. *V Congreso de Fonética Experimental, 2011*. Cáceres.

[22] Nguyen, N., Adda-Decker, M., (eds) 2013. *Méthodes et outils pour l'analyse phonétique des grands corpus oraux*. Paris: Hermès/Lavoisier.

[23] Séguy, J. 1973. Les Atlas linguistiques de la France par régions. *Langue française* 18, 65–90.

[24] Vaux, B., Golder, S. 2003. The Harvard Dialect Survey. Online: http://www4.uwm.edu/FLL/linguistics/dialect/.

[25] Walker, D. C. 2001. *French sound structure*. Calgary: University of Calgary Press.

[26] Walter, H. 1982. *Enquête phonologique et variétés régionales du français*. Paris: Presses Universitaires de France.