Strategies and Challenges for Crowdsourcing Regional Dialect Perception Data for Swiss German and Swiss French

Jean-Philippe Goldman, Simon Clematide, Matthieu Avanzi, Raphael Tandler
University of Zurich, Switzerland
University of Geneva, Switzerland
FNRS/Université catholique de Louvain, Belgium

jeanphilippegoldman@gmail.com

Abstract

Following the dynamics of several recent crowdsourcing projects with the aim of collecting linguistic data, this paper focuses on such a project in the field of Swiss German dialects and Swiss French accents. The main scientific goal of the data collected is to understand people's perception of dialects and accents, and provide a resource for future computational systems such as automatic dialect recognition. A gamified crowdsourcing platform was set up and launched for both main locales of Switzerland: "din dialäkt" ('your dialect') for Swiss German dialects and "ton accent" ('your accent') for Swiss French. The main activity for the participant is to localize preselected audio samples by clicking on a map of Switzerland.

The media was highly interested in the two platforms and many reports appeared in newspapers, television and radio, which increased the public's awareness of the project and thus also the traffic on the page. At this point of the project, 7,500 registered users (beside 30,000 anonymous visitors), have provided 470,000 localizations. By connecting user's results of this localization task to their socio-demographic information, a quantitative analysis of the localization data can reveal which factors play a role in their performance. Preliminary results showed that age and childhood residence influence the how well dialects/accents are recognized.

Nevertheless, quantity does not ensure quality when it comes to data. Crowdsourcing such linguistic data revealed traps to avoid such as scammers, or the participants' quick loss of motivation causing them to click randomly. Such obstacles need to be taken into account when assessing the reliability of data and require a number of preliminary steps before an analysis of the data.

Keywords: Swiss German dialects, French accents, regional variation, cartography, crowdsourcing

1. Introduction

Voices are highly individual and people often wonder where this individuality stems from. Not only we can often guess sociolinguistic attributes like gender, age, and mood from an unknown voice, but also a vast number of characteristics inform us on the speaker's specific regional and social background.

In recent years, several academic projects as well as media initiatives on linguistic variation were launched: see the numerous applications by Leemann et al. (2016) only for English and German dialectal variation; 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 and Bert 2013) for regional variation in the USA; *VerbaAlpina* (Krefeld and Lücke 2014) for dialects spoken in the Alps; and more recently *Donnez Votre Français* (Goldman 2018). They were developed for the web (accessible via computer, tablet or smartphone) or as smartphone applications. A significant number of these projects were initiated in Switzerland, whose variety of languages and unique dialect landscape allows for interesting studies. All of these projects showed the great interest of the public in regiolects, and beyond that, the public's will to understand more about their own voices and the regional linguistic variations of their country.

Following the dynamics of several recent crowdsourcing projects for collecting linguistic data (Cook et al. 2013), the framework presented in this communication focuses on dialects and accents perception. The main goal is to understand people's perception of dialects and accents and provide material for automatic dialect recognition. For this, a gamified crowdsourcing platform was set up and launched in April 2017 for both main locales in Switzerland: "din dialäkt" ('your dialect') for Swiss German dialects and "ton accent" ('your accent') for Swiss French. The main activity is for the participant to localize pre-selected audio samples within Switzerland by clicking on a map. The media was highly interested in the two platforms and many reports appeared in newspapers, television and radio, which increased the public's awareness of the project and thus also the traffic on the page. At this point of the project, 7,500 registered users (beside 30,000 anonymous visitors), have provided 470,000 localizations. By connecting user's results of this localization task to their socio-demographic information, a quantitative analysis of the localization data can reveal which factors play a role in their performance. Preliminary results showed that age and childhood residence influence the how well dialects/accents are recognized. A detailed analysis of each factors' influence will be presented in a paper later this year (Hundt et al. 2018).

Nevertheless, quantity does not ensure quality when it comes to data. Crowdsourcing such linguistic data

revealed traps to avoid such as scammers, or the participants' quick loss of motivation causing them to click randomly. Such obstacles need to be taken into account when assessing the reliability of data and require a number of preliminary steps before an analysis of the data.

We present the linguistic situation in Switzerland, i.e Swiss German dialects and Swiss French accents (Section 2), then describe the audio samples (Section 3) used in the game. In Section 4, we discuss the gamification of the localization task. After describing the participation on the platforms (Section 5), we finally discuss about the drawbacks of a geolocating perception task, the relevance of the socio-demographic information from the users and the reliability of data.

2. Swiss linguistic landscape

In Switzerland, standard German is the official language of 4.5 millions of people (which represents 65% of the population), spread across almost three quarters of the country. French is the standard language of 25% of the population (1.8 millions), gathered on the very western part of the country, in an area called Romandy. Italian and Romansh, the two other official languages of the Confederation, constitute linguistic minorities: they are spoken in small areas in the south and in the south-west of the country.

From a linguistic point of view, varieties of German spoken in Switzerland are noticeably different from standard German: standard German speakers often encounter difficulties when trying to understand Swiss German speakers when the latter communicate in their dialect. Practically, the two linguistic system are in a situation of diglossia: standard German is preferentially used for formal contexts (in particular for writing), while in everyday oral communication, Swiss German dialects are spoken (Ferguson 1959). In addition, it is important to underline that Swiss German is not one uniform language, but consists of a variety of mutually intelligible dialects that differ in terms of lexicon, syntax, morphology and intonation (Siebenhaar 1997: 30). For instance, morphology differs between northern and western dialects, as the former have one or two desinences for the plural of the verb, while the latter have three. Another example for variation in pronunciation would be the presence or absence of diphthongization, etymological [k] evolution, etc., see Sprachatlas der Deutschen Schweiz (1962-2003). This linguistic variation depends, among other factors, on the geographic distance that separate the dialects at stake. While the isoglosses between dialect features draw a highly complex picture of the dialect continuum (visible e.g. in Hotzenköcherle 1984) that allow for a multitude of ways to group small-scale dialects into dialect regions, the general public usually thinks about dialects in terms of political canton boundaries (Siebenhaar 1997: 30).

In the Romandy, the linguistic situation is quite different. Ancestral Gallo-Romance dialects herited from Latin were replaced by French at different stages of history (the process started in the 15th century), and those earlier dialects are barely spoken anymore. Nevertheless, they have left some traces in the French spoken within this area, especially regarding lexicon and pronunciation, which render Swiss French a quite well-identified regiolect when compared with other varieties of European French (Northern/Standard French, Belgian French, Southern French, etc. see Avanzi and Boula-de-Mareüil, 2017, and references therein). In contrast to the aforementioned difficulties for a German to understand a Swiss-German speaking dialect, there are no such obstacles for Standard and Swiss French speakers: a Parisian will easily understand a Swiss French speaker despite his accent, since the two varieties differ only in minor points. Within the Romandy, it is common practice to distinguish Swiss French varieties based on the canton where they are spoken, even if some people will claim that they can make finer distinctions between speakers from different places within the same canton. Lexical cues are usually mentioned by Swiss people when asked which feature they based a localization on, but usually it is the accent (pronunciation variants) that allows one to identify the region of origin of a given Swiss French speaker.

3. Audio data

The auditory input in the games was drawn from several different corpora for French and German. For the French game, all the samples were drawn from the OFROM corpus, which was recorded and transcribed by the University of Neuchatel between 2008 and 2014 (Avanzi et al. 2015). The samples for the German game were extracted mainly from three corpora: the SDS Phonogramme (Phonogrammarchiv Zürich 2001), Archimob (Vereinigung Archimob 2000; Samardžić et al. 2016) and Stimmen der Schweiz (Glaser/Loporcaro 2012). Due to a lack of speakers in the cantons of Appenzell Innerrhoden, Grisons, Aargau and Fribourg in the corpora mentioned above, a few samples were extracted from other sources (see acknowledgments section).

Here is a transcribed example for each language:

- a male French-speaker from Vaud, born in 1925 (87 years-old during the recording), retired, former wine-producer:

ben maintenant ils | ils remplissent leur cageot donc fin euh enfin leur caisse | c'est toujours les caisses | mais | ils vont plus laisser y a y a des | les dames qui vendangent | pis y a des jeunes qui portent la caisse

well now they | they fill their crate | so | well their boxes| it's always boxes | but | they won't let, there is | there is some | the ladies who harvest | and there are young people who carry the box

- a male German-speaker from Appenzell Innerrhoden born in 1957 (60 years-old during the recording), butcher

Aso gfloge simmer no nü meng mal | mier sind, äh, mal, ääh | vor sechs Jahr simmer uf Gran Canaria | mit de Gove | d'Meitli hend gmeint sie wettid öu e chli as Meer id Ferie | de simmer uf Gran Canaria | und denn hemmer gseit, da hemmer jetzt gseh, etz müemmer nöd unbedingt

So we haven't flown many times | we went, uh, once, uuh | to Gran Canaria six years ago, with the kids | the girls said they wanted to go to the beach a little during holidays | so we went to Gran Canaria | and then we said, we've seen it now, now we don't have to [anymore] necessarily

Certain constraints on the selection of the samples were introduced. In terms of duration, the samples had to be between 10 and 20 seconds. Content-wise, samples could neither contain any personal names, in order to ensure the anonymity of the speakers, nor any geographical clues that would hint towards the speaker's location. Furthermore, the aim was to exclude any hints pointing towards regional culture in the content, for instance in form of regional legends or politics.

Among the pre-selection of 3 to 9 samples that were extracted from the audio-files in accordance with the aforementioned constraints, one sample was chosen based on the auditory quality (not too many long breaks; no great divergence of loudness; no distracting sounds) and whether the content would be interesting to the players. No speaker appears more than once, in order to ensure that users do not perform better by remembering a voice.

Unlike other studies in the field of perceptual dialectology, such as Baker et al. (2009), linguistic features were not taken into account for the selection of the samples. In contrast to the corpus of the Baker study that consisted of speech based on a pre-written dialogue, the corpora for this project contained mostly free speech. Considering the variety of the many recordings, controlling the salient features would not have been possible.

As a consequence of not controlling the linguistic features within the single samples, the difficulty of locating a specific sample might also have differed depending on the amount of salient features of a dialect that appeared in a sample, especially in the case of lexical or certain phonological cues. However, as all the samples are longer than 10 seconds, it is very unlikely that no salient features would be included in any of the samples. Furthermore, the difference of linguistic features can also be used as a fruitful basis for a study on salience by comparing the recognition rate of samples of the same dialect that include or exclude a certain feature.

There are a number of factors concerning the samples that need to be taken into account in the analysis. As mentioned above, the samples in the game had to be drawn from a variety of sources due to the lack of an extensive spoken corpus of contemporary Swiss German. Two of these sources, the SDS Phonogramme and the Stimmen der Schweiz, which form a substantial part of the samples, are recordings made approximately between 1940 and 1960. For this reason, the audio quality of some of the older samples is not of the same standard as the newer ones. Furthermore, the speakers recorded in these two projects were older people whose dialects nowadays are not spoken in their way anymore. This means that not only knowledge of current, but also of older dialects determined the performance of the player. As a consequence, age might become a more important factor, as the older players might still have had contact with those older stages of the dialect.

4. Gamification

In order to keep up the motivation of the participant, we had go beyond his/her interest into regional variation and we turned the localisation task into a facilitated and gamified activity by several means:

- Instead of offering a long-haul series of samples to localise on a map (150 samples in French and 114 samples in Swiss-German), we **organized the game in rounds** of 10 samples. Dividing this task in sub-tasks is much more attractive.
- Another way to motivate users is to offer them simple tasks at first then lead them **progressively more difficult tasks**. In that sense, we split our game in two modes (see Table 1):
  - In the so-called **easy mode** ("einsteigen" task in the German game, "amateur" for the French one), the player has to choose one canton among a pre-selection of cantons. For each of the Swiss-German samples, a different set of 5 cantons (out of 20 existing ones) was selected. For the Swiss French game, all the cantons were selectable, as there are only seven. Scoring is binary, i.e either the player scores 100 points or nothing (see Figure 1).
  - In the **expert mode**, the player has to click on an exact location within the Swiss map. The score depends on the distance between the user's click and the exact location of the audio sample. A *grace area* of 10 kilometers in radius around the sample's location gives the user the maximum number of points (see Figure 2).
  - As the "klettern" task with its unrestricted localization could be rather challenging and frustrating, the users first had to complete the "einsteigen" task, where most players would be more successful.
- Newcomers discover the game with a **trial round** of 3 samples in each mode (easy and expert), then they are incented to register with email address, username

and password only. This trial round is also considered as a training stage to ensure that they have understood the mechanics of the game before their performance is recorded. In order to avoid that our participants do not complete the registration process, we delay the presentation of the form with socio-linguistic questions (after few answers), we also ask these various informations one by one after each round. Of course, they are offered to fill up the socio-linguistic form at once.

- We also **gamified** the tasks with points and set up a leaderboard where users can see their own ranking as well as the top scorers.
- Finally, some efforts were done to facilitate the **user experience** (e.g simplified instructions, blinking buttons to assist navigation)

	Easy mode	Expert mode
de	trial round of 3 samples 3 rounds of 7 samples total = 24 samples	trial round of 3 samples 9 rounds of 10 samples total = 93 samples
fr	trial round of 3 samples 2 rounds of 7 samples total = 17 samples	trial round of 3 samples 13 rounds of 10 samples total = 133 samples

Table.1 Nb. of rounds and samples per language and mode

	regis. users	childhood residence	birth decade	gender
de	3705	62%	55%	49 %
fr	2519	50 %	42 %	36%

Table 2. Proportion of sociolinguistic information provided by the participants.

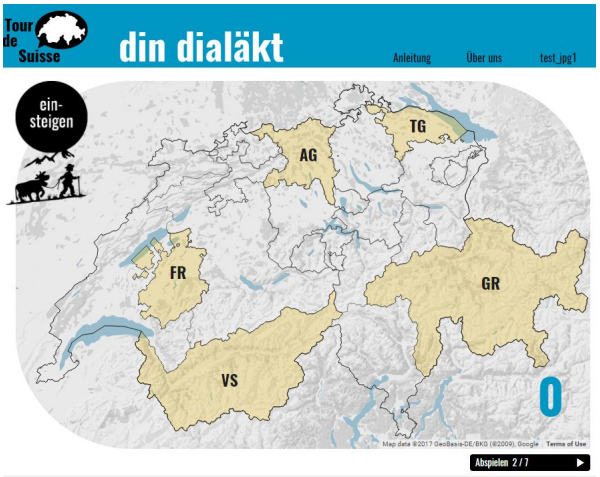

Figure 1. Easy mode for Swiss German (5 possible answers)

Figure 2. Expert mode for Swiss German (after the participant clicked onto the map)

5. Crowdsourced results

After few months online, the platform had a great success and allowed us to gather a lot of data. About 7'500 users have registered. The birth decade was given by 43% of them and their childhood residence by 56 % of them. In the end, about 470'000 tasks on map were achieved. The logs showed that a number of players quickly lost interest (Figure 3 and 4) and that the proportion of them who went through all the rounds is rather limited (19% for DE and 6% for FR, see the rightmost bins). Nevertheless, preliminary results showed interesting effect of age and childhood residence to be further evaluated in final version of this paper.

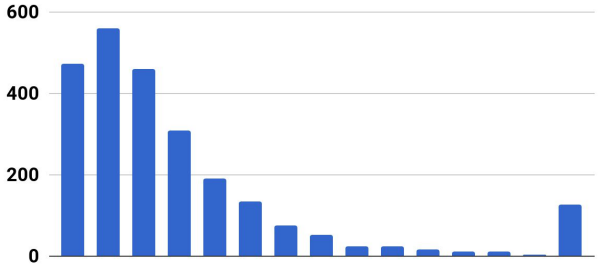

Figure 3. Number of French users as number of completed rounds (from 1 to 15 rounds)

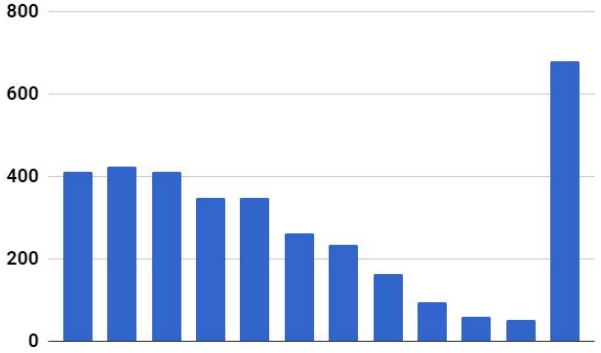

Figure 4. Number of Swiss German users as number of completed rounds (from 1 to 12 rounds)

Here are some map results of the expert game for German and French (Figures 5 to 8), where the individual answers

(red dots) are superimposed with "heat-map" modeling and the ground truth (green cross):

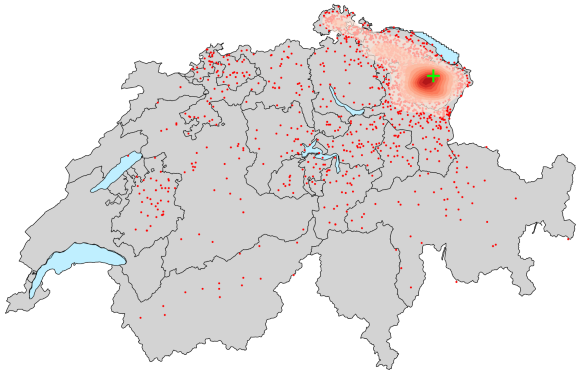

Figure 5: Correct global identification of a German-speaker from St.Gallen (task #494)

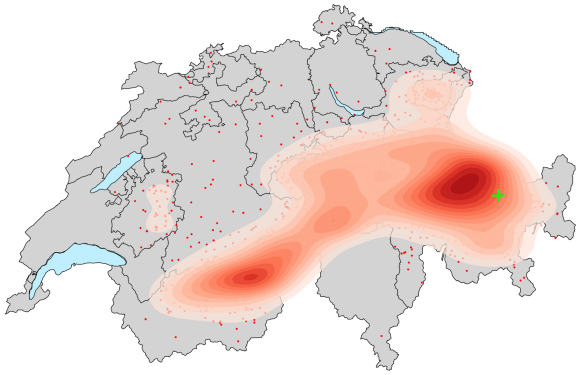

Figure 6: Bi-modal recognition of a German-speaker from Grisons (task #492)

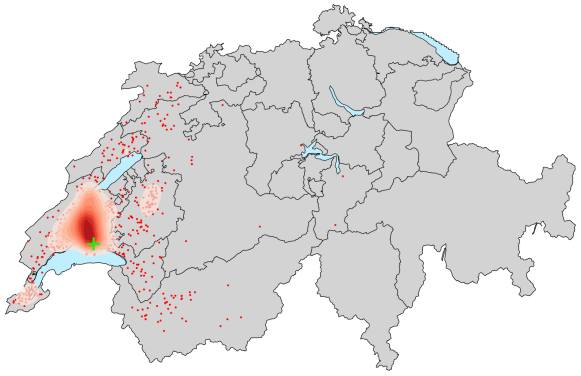

Figure 7: Correct global identification of a French-speaker from Lausanne (task #331)

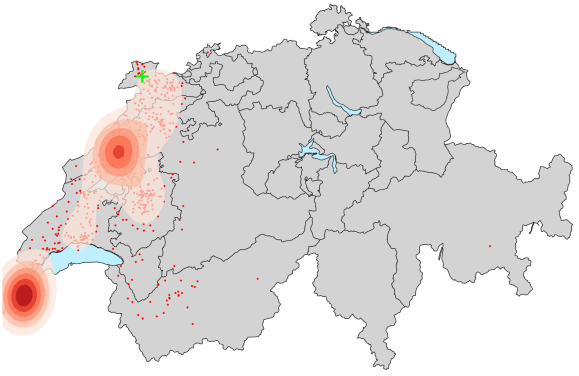

Figure 8: Bi-modal recognition of a French-speaker from Jura (task #332). The two modes are located in Geneva (lower group) and Neuchâtel (upper group).

6. Discussion

On the basis of this experiment, various questions can be addressed and debated:

1. **Danger of geolocating: are we testing the geographical competence and/or dialect perception competence?** The selection of geographic information displayed on the map is crucial (administrative boundaries, cities, main roads, water bodies) as they may influence the answer regarding both, the geographical knowledge and the dialectal landscape representation of the player. Beside, dialectal areas may not always coincide with administrative ones. We stuck to these latter ones in the easy game, as people think in terms of those cantonal boundaries, and there was no other simple alternative that would have been a lot more exact that those boundaries. Finally, geographical information represented as points, for instance, cities, can be attractors and introduce a bias in the spatial distribution of the answers.
2. **Getting the relevant socio-demographic information from the users: what should we ask as personal info?** Asking too much information from the player about his/her linguistic history may be frightening or boring. At the same time, information that is not precise enough might render an analysis more difficult. Thus, one has to balance between not getting enough information if it is too personal, and not getting the relevant data if it is too general.
3. **Population of informants: do we get even distribution of participants?** Such online surveys, even relayed by popular media, ineluctably bring up a younger and more connected population. Another consequence is that towns are overrepresented compared to countryside areas. This problem of an uneven distribution of players in terms of age and location needs to be addressed.
4. **Reliability of data: is everyone who takes part of the game doing it seriously?** Closer observations of the data showed some types of players that should be

certainly discarded, but the criteria of removal need to be set carefully; for example, users that perform exceptionally badly (which could be unmotivated players, children or people with a very low knowledge of the dialectal variety, or elderly people struggling with the technological aspect), or users with a very high score (who could be cheaters aiming to get points and a high ranking in the leaderboard, or linguistic experts).

The aim of the paper being a presentation of the crowd-sourcing framework, some complementary data and further analyses of the results are necessary, assessing in more details the results as well as the benefits and drawbacks of such crowdsourced linguistic data.

7. Acknowledgements

This research is supported by the Swiss National Science Foundation as Agora project n°164811. We thank the GISLab team from URPP Language and Space (Research Priority Program) and the SIVIC team (Scientific Visualisation and Visual Communication) of the University of Zurich, Switzerland.
Audio source: for the samples from Grisons, we thank Hanna Ruch who recorded speakers from there for her project "Akkommodation in Dialektkontaktsituationen" financed by the URPP Language and Space; for the ones from Fribourg we thank Ivan Schmutz from the Sensler Museum in Tafers; for the ones from Aargau we thank Nicole Studer-Joho and Dieter Studer-Joho. The samples from Appenzell-Innerrhoden were recorded by Raphael Tandler from the dindialaekt-team.

8. References

Avanzi, M., Béguelin, M.-J. Diémoz, F. (2015). « OFROM – corpus oral de français de Suisse romande, v. 2.2 », Ms, Université de Neuchâtel, http://www.unine.ch/ofrom

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

Avanzi, M. and Boula-de-Mareuil, Ph. (2017). Perceptual identification of regional French accents in (northern) France, Belgium and Switzerland, *Journal of Linguistic Geography*, 5(1):17-40.

Baker, Wendy, David Eddington and Lyndsey Nay (2009). Dialect identification: The effects of region of origin and amount of experience. *American Speech*, 84(1): 48-71.

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.

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.

Ferguson, Charles A. 1959. "Diglossia". *Word* 15: 324–340.

Glaser, Elvira and Michele Loporcaro (eds.) (2001). *Stimmen der Schweiz*. Frauenfeld: Huber. 1 book and 2 CDs.

Goldman J.Ph. et al (2018) Crowdsourcing Regional Variables and Automatic Geolocalisation of Speakers of European French, LREC Conference, Miyasaki, Japan.

Hotzenköcherle, Rudolf. 1984. *Die Sprachlandschaften der Schweiz*. Ed. N. Bigler, R. Schläpfer and R. Börlin. Aarau/Frankfurt: Sauerländer.

Hundt, Marianne, Raphael Tandler and Karina Frick. Localization and perception of Swiss German dialects: comparing performance and judgement data. Forthcoming 2018.

Krefeld, T., S. Lücke (2014). VerbaAlpina - Der alpine Kulturraum im Spiegel seiner Mehrsprachigkeit. In Ladinia XXXVIII, 189-211.

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

Phonogrammarchiv Zürich (eds.) (2001). *SDS-Phonogramme*. 8 CDs + 4 CD booklets.

Samardžić, Tanja, Yves Scherrer and Elvira Glaser (2016). *ArchiMob – A Corpus of Spoken Swiss German*. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). Online.

Siebenhaar, Beat and Aflred Wyler. 1997. *Dialekt und Hochsprache in der deutschsprachigen Schweiz. Zürich: Pro Helvetia.*

Sprachatlas der Deutschen Schweiz. (2017). Bern (I-IV), Basel:Francke (VII-VIII).

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

Vereinigung Archimob (eds.) (2000). *Archive der Zeit des zweiten Weltkrieges in der Schweiz: 383 deutschsprachige Zeugen*.