Big Data for analyses of small-scale regional variation: A case study on sound change in Swiss German

Adrian Leemann[1], *Marie-José Kolly*[2]

[1]Phonetics Lab., Department of Theoretical and Applied Linguistics, University of Cambridge
[2]Phonetics Lab., Department of Comparative Linguistics, University of Zurich
al764@cam.ac.uk, marie-jose.kolly@uzh.ch

Abstract

In this case study we examine sound change of *Altoberdeutsch* <iu> in Swiss German dialects. We used contemporary dialect data from nearly 60,000 speakers – collected with the smartphone app *Dialäkt Äpp* – and compared it to historical *Atlas* data from the 1950s. Results revealed hierarchical and contra-hierarchical diffusion patterns for some dialectal variants, while other variants remained virtually unchanged over the course of seven decades. We further report change in apparent time, with older speakers using traditional variants more frequently than younger speakers. Using this case study as a model, future work using the *Dialäkt Äpp* corpus will reveal patterns of feature diffusion and dialect leveling on a larger scale.

Index Terms: sound change; dialect leveling; crowdsourcing; dialectology; Swiss German

1. Introduction

The most recent large-scale study on Swiss German (hereafter SwG) dialects – the *Sprachatlas der Deutschen Schweiz* (*Atlas* for short, [1]) – dates back 60–70 years. It documents the linguistic situation of German-speaking Switzerland in the first half of the 20th century for 566 localities. Anecdotal evidence and previous, mostly small-scale studies, revealed that dialects have changed considerably since then. Yet our understanding of how dialects have changed on a regional level remains patchy. In this contribution, we will contribute to fill this gap with a case study using Big Data that was crowdsourced with the free iOS app *Dialäkt Äpp* (*DÄ* for short, [2]). *DÄ*'s main function is the prediction of the user's dialect [3]. For 16 variables, users select their dialect variant from a drop-down menu. For the variable *Bett* 'bed', for example, they choose from the variants [bɛt] (as used in Western SwG) or [bet] (Eastern SwG). At the end of the quiz, *DÄ* guesses which dialect the user speaks. Underlying this prediction are 16 maps from the *Atlas*. Following dialect prediction, users can evaluate the result and indicate their actual dialect. With this information, the 16 variables can be assessed for language change (*Atlas* vs. *DÄ* data). A first pilot revealed global patterns of language change in SwG [4]. The large bulk of this corpus, however, is yet to be analyzed, especially with regard to in depth, small-scale analyses of regional variation and change. The objective of the present proof of concept study is the analysis of small-scale, regional diachronic variation in SwG dialects in the variable *Altoberdeutsch* <iu>.

1.1. Previous studies

Only a few studies have examined how SwG dialects have changed since the *Atlas*. [5] and [6] reported change in the lexicon. The latter found convergence tendencies towards Standard German and showed that younger speakers deviated from the *Atlas* more than older speakers in lexical features. Similarly, [5] conducted an online survey with 9000 participants. Based on this study, [7] as well as [8] presented dialect maps that corroborate tendencies of leveling in the lexicon for some of the variables examined. For morphosyntax, [9] found that only little change has occurred for the constructions examined. A number of studies have further reported sound change over the past decades, such as [10], [11] and [12]. The two latter investigations documented significant change for Aarau. On the whole, variants that were documented in the *Atlas* are still in use in Aarau, yet they co-exist alongside additional, more frequently used variants. /l/-vocalization in particular has received much attention in the literature. A number of studies report the spread of vocalized /l/ to regions not previously attested as vocalizing in the *Atlas*: towards Luzern [13, 14], Fribourg [15], Central Switzerland [16], and the Bernese Oberland [17, 18]. Our own research revealed change between the *Atlas* and today: a recently conducted study applying a rapid anonymous survey framework [19] indicates that /l/-vocalization has spread in a southerly, westerly, and central direction within German-speaking Switzerland. [4], using the *DÄ* corpus, revealed that phonetic variables demonstrated most agreement with the *Atlas* (67%), followed by the morphological (59%), and the lexical variable (53%). Until today, however, we have not investigated small-scale regional patterns of language change to the level of detail required, using *DÄ* data.

1.2. Research questions

In the present contribution, we provide a case study for small-scale, diachronic analyses of one variable, *Altoberdeutsch* <iu>, which – in most cases – stems from (Proto-)Germanic <eu> [20]. Both [20] and [21] claim <eu> to be one of the most complex variables with regard to how the sound has changed over time. (Proto-)Germanic <eu> developed towards *Altoberdeutsch* (the southern varieties of Old High German) <iu> in words such as *tief* 'deep' or *Fliege* 'fly' while varieties further north featured <oi> [20]. In Middle High German <iu> began changing into three spatially distributed variants in Switzerland: (i) a Northeastern variant <üü>, (ii) a Northwestern diphthongized variant <ie>, and (iii) a Southwestern group of variants which underlyingly trace back to <öü> (cf. Figure 3; [21]).

2. Methods

Section 2.1. describes the prediction and evaluation functionality of *DÄ* that enables analyses of language change; in 2.2. we describe the users and localities of the *DÄ* corpus, and 2.3. presents the distribution of *tief* variants as represented in the *Atlas* – the reference point for analyses of sound change.

2.1. Dialäkt Äpp & Procedures

DÄ's main function is the prediction of the user's dialect, which is based on 16 discriminative maps from the *Atlas*. The app prompts users to select their pronunciation variant from a list of each of the 16 variables by tapping on the screen. Given that SwG does not have a standard writing system, variants are spelled close to pronunciation – and feature additional IPA transcriptions where necessary. All variants are accompanied with sounds for users to listen to; see the prompt for *tief*, Figure 1.

Figure 1: *tief* and its dialectal variants for users to select.

When users arrive at the end of the quiz, the app presents a list of five localities – out of 550 adapted from the *Atlas* (16 / 566 original localities have merged, [11]) – that best corresponds to the user's dialect. Users can then evaluate the predicted dialect (Figure 2, left). In case of an accurate prediction, they type in age and gender and send off the data anonymously (Figure 2, center left). In case of an incorrect prediction, they indicate their dialect by choosing from a list of localities (Figure 2, center right), which correspond to those used for the dialect prediction; users select age and gender and send off their information (Figure 2, right).

Figure 2: Evaluation of dialect prediction by users.

We then compare the users' values to those in the *Atlas*. *DÄ* predicted Bern for the fictive user of Figure 2 (center left). If s/he in fact speaks the Bern dialect, s/he would enter age and gender, and send off the data. The 16 variants for Bern as indicated in the *Atlas* are then compared to the values entered by the user. In the case shown here, it is likely that there is little discrepancy between the *Atlas'* and contemporary values, since *DÄ* predicted the correct locality. If, however, this user claimed to speak the dialect of Burgdorf, s/he would indicate this (as shown in Figure 2, center right and right) and send off the data. In this case we obtain a greater difference between the *Atlas'* and the contemporary data, which means that the dialect of this speaker from Burgdorf has become more like the dialect of Bern.

2.2. *DÄ* corpus

The corpus consists of data from 58,923 users from effectively all localities in the *Atlas*. Only three *Atlas* localities were not represented in the DÄ corpus: Mutten (Grisons), Obergoms (Valais), and Sternenberg (Zurich). For all other localities, there was at least one respondent, with a median of 48 respondents per locality. 42% of the users were females, 58% males. On average, users were 31.5 years old (MD=27; SD=15.5). 30% of the users were predicted in the correct locality, and 65% in the right canton [4].

2.3. Reference material

For the sake of showing results on analyses of sound change (cf. 3.2. & 3.3.), Figure 3 shows the *Atlas* variants.

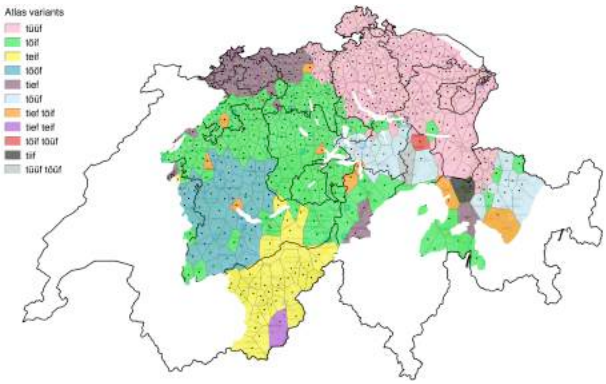

Figure 3: Variants of *Altoberdeutsch* <iu> shown in the *Atlas*.

Northeastern SwG had <üü>, Northwestern SwG <ie>, and in the multi-variant region (cf. 1.2) we find <öi> in the Western Midlands and Central Switzerland, and <öö> in much of the Southwest [21]. The *Atlas* further indicated large areas of <öü> in parts of Graubünden and in Central Switzerland. Fribourg and the Southwestern part of Bern are characterized by monopthongized <öö>. <ei> was reported in Southeastern Bern and in Valais. A pocket in Uri featured <ie>, which otherwise is dominant in the Northeast. Some of the variants shown in the *Atlas* were categorized for *DÄ* (e.g. <täüf> was included in <töüf>), and a number of localities demonstrated two variants (see Figure 5). Space prevents a comprehensive review of this categorization procedure; it was conducted using plausible historical linguistic rationales.

3. Results

3.1. Number of respondents

Figure 4 shows the number of respondents per locality, broken into ten natural classes (Jenks). We used *Voronoi* polygons for each locality (ten buffer) in Figures 3–6. Midland localities (e.g. Zurich N=3119, Bern N=2736, Basel N=1842) show the

highest number of respondents; Alpine localities, on the contrary, frequently feature 1–40 respondents only (e.g. Habkern (Bern) N=11, Betten (Valais) N=9, Sisikon (Uri) N=16).

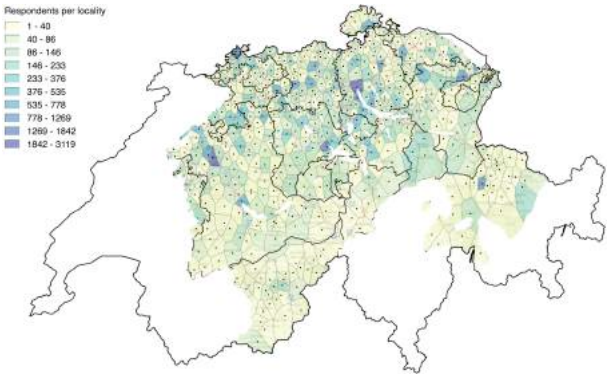

Figure 4: Number of respondents by locality.

3.2. Agreement with Atlas

Figure 5 shows the *DÄ–Atlas* agreement scores – 0 (red) showing no agreement, 1 (green) showing full agreement.

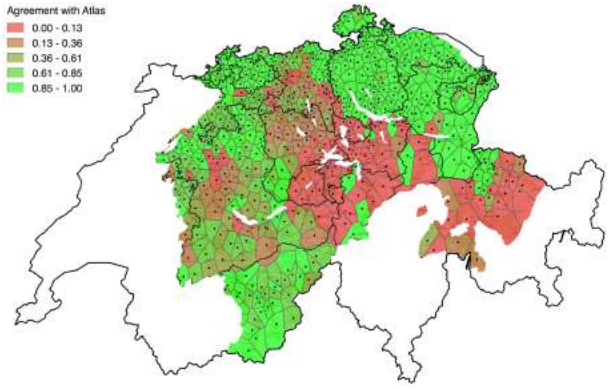

Figure 5: Agreement of *DÄ* variants with *Atlas* variants.

Much of Northeastern Switzerland reveals high agreement scores (green), e.g. the cantons of Zurich (cantonal mean M=.94), Thurgau (M=.96), St. Gallen (M=.92), as well as both Appenzell cantons (AI, M=.75; AR, M=.93). Much of Central Switzerland (e.g. Nidwalden, M=.015), some localities in the canton of Bern (M=.61), many localities of the cantons of Aargau (M=.54) and Graubünden (M=.35) reveal high disagreement scores.

3.3. *DÄ* variants

Figure 6 illustrates the variants indicated in *DÄ*. The broad geographical patterns attested in [20], [21], and in the *Atlas* (cf. Figure 3) remain largely intact: <üü> in the Northeast, <ie> in the Northwest, and a multi-variant region in the Southwest. The isoglosses of <ie> in the Northwest appear to have remained stable; <üü> has gained considerable terrain, spreading towards the Southwest, where it replaced <öü> in Graubünden, Glarus and Schwyz, and pushed aside <öi> in most of Aargau and Luzern. The geographical distribution of unrounded <ei>, mostly present in the Wallis, remained stable. Quite strikingly, <ie> – a typical feature of Basel German, but also, to a small extent, found in Uri (cf. Figure 3) – has diffused towards numerous Central Swiss localities, replacing <öi> in the Cantons of Uri, and Unterwalden. This phenomenon is also illustrated in Figure 5 where many of the Central Swiss localities are colored in red (indicating much disagreement with the *Atlas*). In Southern Bern <öi> replaced the monophthongized variant <öö>. For some localities, *DÄ* data included the same proportion of speakers for two different variants. <öi> and <öö>, for example, were equally reported in one locality nestled between the <öi> / <öö> isogloss, see Figure 6 (dark blue).

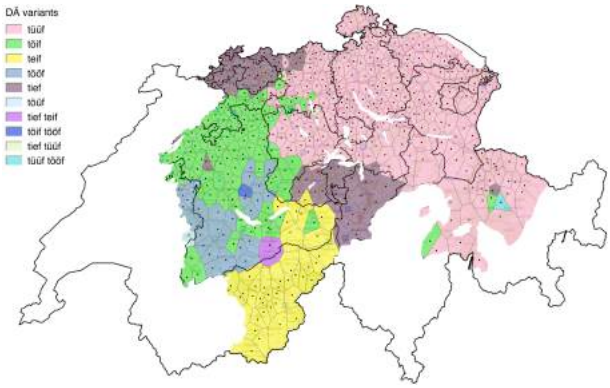

Figure 6: Variants of *Altoberdeutsch* <iu> as used in *DÄ*.

3.4. Change in apparent time

Figure 7 shows the *Atlas* agreement scores by age group. We divided the speakers into five natural breaks (Jenks) according to their age: 10–22, 23–32, 33–43, 44–57, 58–90. To test for an effect of *age*, we ran a GLM that included *sex*, *age* and *dialect* as factors (α=.05). A standard likelihood ratio test revealed a significant effect of *age* ($\chi^2(4)$=-481.07, p<.0001). Figure 7 shows that the oldest group (purple) used the variants indexed in the *Atlas* the most (high agreement), the youngest group (red) the least (low agreement).

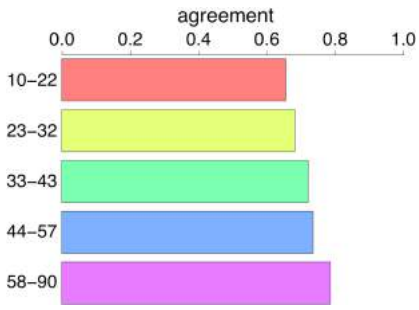

Figure 7: Proportion of *DÄ* speakers who indicated the same variant as indicated in the *Atlas*.

For the youngest age group, we found that 66% of answers were identical with the *Atlas* (red), followed by 68% (yellow), 72% (green), and 74% (blue). The oldest group (purple) demonstrated the highest *Atlas* agreement scores with 79% (purple).

4. Discussion

4.1. Regional variation and change in apparent time

A relatively recent theme of work in dialectology is leveling: the loss of minority dialectal variants and regional convergence towards majority features (cf. [6, 7, 8]). We find such leveling tendencies for *tief*: the Northeastern variant <üü> has spread considerably in southerly, westerly, and southeasterly direction – resulting in a convergence of traditional forms used in the 1950s towards the majority form <üü>. Reasons for this change are multifactorial and allow only for speculations on our part. The change in Central Switzerland may have to do with a substantial figure of the students from Schwyz and Glarus commuting to Zurich for training [23]. This spread may be evidence for an example of hierarchical diffusion, where the majority form spreads to increasingly smaller settlements [24]. Secondly, the diffusion of <üü> in southeasterly direction may be explained with residents from rural Grisons settlements commuting to work in Chur [25, 23], a city that serves as a transportation and cultural hub of the area and which – most importantly – was reported as using <üü> in the *Atlas* (see Figure 3). In addition to this potential diffusion from within the canton, Southeastern Switzerland is a popular summer and winter holiday destination for residents from the canton of Zurich [26]. Thirdly, we find a diffusion of <üü> in a westerly direction towards the canton of Aargau, pushing back local variants such <öi> and <ie> in Aarau, Lenzburg, Bremgarten, and Muri. Only the Bernese Aargau (Zofingen district) remained relatively stable in the southern periphery. Previous studies have shown that this canton in particular has proven to be in flux and in a zone of dialectal instability, nestled between the two major linguistic radii of Bern and Zurich [11].

Looking at the change of other variants, the monophthongized variant <öö> has lost substantial ground in Western German-speaking Switzerland and seems to be becoming replaced by <öi>, possibly under the influence of the linguistic radius of the urban regions of Bern (cf. [19]). What stands out synchronically, however, is the Bern city-specific variant <ie> in contemporary data, despite being surrounded by localities reporting <öi>. This Bern city-specific variant does not seem to spread to its urban regions. Another noticeable process of diffusion appears in Central Switzerland, where <ie> has diffused substantially towards both Ob- and Nidwalden, as well as to virtually all localities elicited in the canton of Uri. Presumably this is an example of contra-hierarchical diffusion (cf. [27]), where changes spread from the rural region – e.g. Andermatt (Uri) – to smaller towns and finally to larger towns (e.g. Sarnen, Stans).

The *age* effect reported in 3.4. fits in nicely with our intuitions on sound change in apparent time in SwG: older speakers show different speech patterns than younger speakers, which can be understood as an instance of sound change taking place in our sample, as the older speakers' variants agree with those reported in the *Atlas* more often. This trend may, however, also be an artifact of the large data set we are using (cf. Kilgarriff 2005).

4.2. Methodological caveats

There are methodological caveats that need to be kept in mind when we compare data that has been crowdsourced or collected with traditional dialectological methods (cf. [4]). In the *Atlas*, researchers elicited data directly from the speakers. For *DÄ*, there was no researcher present – giving subjects substantial freedom of interpreting the instructions given. For the *Atlas*, speaker selection criteria were stringent – as was typical for dialectology at the time [28]; in *DÄ*, users had different linguistic backgrounds, educational levels, and mobility habits. The two databases further differ in the distribution of speaker age: *Atlas'* subjects ranged between 51 and 80 [29]; in *DÄ* the median age is 27 [4]. Further factors that contribute to noise in the data are the users' self-declaration of dialects when evaluating the result, where users may have imitated a 'model', perhaps more nostalgic form of a dialect when doing so; potential multiple submissions [30], and potential biases stemming from the user interface (variants presented at the top of the drop-down menu may have been clicked more often than those at the bottom). Despite this noise, previous research has shown that traditional dialectological methods reveal very similar diffusion patterns to those found through app crowdsourcing [4].

5. Conclusions

We presented a case study of how Big Data, crowdsourced through a smartphone app, can be used to study small-scale regional diachronic variation. From a methodological viewpoint, this dataset provides a novel way of studying language change due to the new sampling technique: dialectological methodology embodies a notion of the 'authentic' speaker; it has been biased towards population groups associated with maintaining the most distinctive regional varieties, i.e. NORMs [28, 29] or speakers of the 'vernacular' [30]. By changing data collection methods and giving up control over sampling, our approach avoids these biases. This approach is not meant to replace existing techniques for the collection of dialectological data, but simply wishes to highlight the power and added value of crowdsourced Big Data as a way of complementing established methods. Using this case study as a model, future studies using this corpus will reveal in greater detail which areas have undergone most change and which variants have spread or been replaced in the past 60–70 years.

6. Acknowledgements

We thank Daniel Wanitsch for server-side technical assistance and 65 backers who made *DÄ* possible through crowdfunding.

7. References

[1] Sprachatlas der deutschen Schweiz. (1962–2003). Bern: Francke (Vols. 1–6), Basel: Francke (Vols. 7, 8).

[2] Leemann, A., & Kolly, M.J. (2013). Dialäkt Äpp. https://itunes.apple.com/ch/app/dialakt-app/id606559705?mt=8 (accessed 30.06.2016).

[3] Kolly, M.-J., Leemann, A. (2015). Dialäkt Äpp: communicating dialectology to the public – crowdsourcing dialects from the public, in: Leemann, A., Kolly, M.-J., Schmid, S., Dellwo, V. (Eds.), Trends in Phonetics and Phonology. Studies from German-speaking Europe (pp. 271–285). Bern etc.: Lang.

[4] Leemann, A., Kolly, M.-J., Purves, R., Britain, D., Glaser, E. (2016). Crowdsourcing language change with smartphone apps. PLoS ONE 11/1: e0143060.

[5] Glaser, E. (2008). Der Wortschatz des Schweizerdeutschen. http://www.ds.uzh.ch/Forschung/Projekte/Schweizer_Dialekte/index.php (accessed 30.06.2016).

[6] Juska-Bacher, B. (2010). Wortgeographischer Wandel im Schweizerdeutschen. Sommersprossen, Küchenzwiebel und Schmetterling 70 Jahre nach dem SDS. Linguistik online, 42, 19–42.
[7] Christen, H., Glaser, E., Friedli, M. (Eds.). (2015). Kleiner Sprachatlas der Deutschen Schweiz. 6th ed. Huber: Frauenfeld.
[8] Glaser, E. (2016). Weiterführung der Online-Umfrage 2008. http://www.ksds.uzh.ch/de/onlineumfrage2008.html
[9] Glaser, E. (2014). Wandel und Variation in der Morphosyntax der schweizerdeutschen Dialekte. Taal en Tongval, 66(1), 21–64.
[10] Christen, H. (1998). Dialekt im Alltag: Eine empirische Untersuchung zur lokalen Komponente heutiger schweizerdeutscher Varietäten. Tübingen: Niemeyer.
[11] Siebenhaar, B. (2000). Sprachvariation, Sprachwandel und Einstellung. Der Dialekt der Stadt Aarau in der Labilitätszone zwischen Zürcher und Berner Mundartraum. Stuttgart: Steiner (Zeitschrift für Dialektologie und Linguistik, Beihefte 108).
[12] Siebenhaar, B. (2002). Dialektwandel und Einstellung – Das Beispiel der Aarauer Stadtmundart, in: Berns, J., van Marle, J. (Eds.), Present-day Dialectology: Problems and Findings (pp. 313–332). Berlin, New York: Mouton de Gruyter (Trends in Linguistics 137).
[13] Haas, W. (1973). Zur l-Vokalisierung im westlichen Schweizerdeutschen, in: Bausinger, H. (Ed.), Dialekt als Sprachbarriere: Ergebnisbericht einer Tagung zur alemannischen Dialektforschung (pp. 63–70). Tübingen: Tübinger Vereinigung für Volkskunde.
[14] Christen, H. (1988). Sprachliche Variation in der deutschsprachigen Schweiz. Dargestellt am Beispiel der l-Vokalisierung in der Gemeinde Knutwil und in der Stadt Luzern. Stuttgart: Steiner (=Zeitschrift für Dialektologie und Linguistik. Beiheft 58).
[15] Piller, A. (1997). Sprachwandel im Sensebezirk dargestellt am Beispiel der /l/-Vokalisierung und der Rundung der Palatalvokale. Licentiate thesis, University of Fribourg.
[16] Christen, H. (2001). Ein Dialektmarker auf Erfolgskurs: Die /l/-Vokalisierung in der deutschsprachigen Schweiz. Zeitschrift für Dialektologie und Linguistik, 1, 16–26.
[17] Flury, A. (2002). l-Vokalisierung und nd-Velarisierung in Spiez: Eine empirische Untersuchung. Licenciate thesis, University of Bern.
[18] Matter, M., Ender, A. (2006). Datenerhebung mit einer Rapid Anonymous Study am Beispiel der l-Vokalisierung. Talk at 4. Tage der Schweizer Linguistik, 20 November 2006, Basel.
[19] Leemann, A., Kolly, M.-J., Werlen, I., Britain, D., Studer-Joho, D. (2014). The diffusion of /l/-vocalization in Swiss German. Language Variation and Change 26(2), 191–218.
[20] Wiesinger, P. (1970). Phonetisch-phonologische Untersuchungen zur Vokalentwicklung in den deutschen Dialekten. Vol. II: Die Diphthonge im Hochdeutschen. Berlin: de Gruyter.
[21] Hotzenköcherle, R. (1986). Dialektstrukturen im Wandel. Gesammelte Aufsätze zur Dialektologie der deutschen Schweiz und der Walsergebiete Oberitaliens. Aarau, Frankfurt am Main, Salzburg: Sauerländer.
[22] BFS=Bundesamt für Statistik (2016). Amtliches Gemeindeverzeichnis der Schweiz. www.bfs.admin.ch (accessed 30.06.2016).
[23] BFS=Bundesamt für Statistik (2015). Pendlermobilität der Schweiz 2013. Neuchâtel.
[24] Bailey, G., Wikle, T., Tillery, J., Sand, L. (1993). Some patterns of linguistic diffusion. Language variation and change, 5(3), 359–390.
[25] Amt für Raumentwicklung Graubünden (2007). Siedlungsbericht Graubünden. Analyse der Siedlungsentwicklung seit 1980. Chur.
[26] Amt für Raumentwicklung Graubünden (2012). Zweitwohnungen in Graubünden. Canobbio.
[27] Wikle, T. (1997). The spatial diffusion of linguistic features in Oklahoma. Proceedings of the Oklahoma Academy of Science 77, 1–15.
[28] Chambers, J. K., Trudgill, P. (1998). Dialectology. Cambridge University Press: Cambridge.
[29] Hotzenköcherle, R. (1984). Die Sprachlandschaften der deutschen Schweiz. Edited by Bigler, N., Schläpfer, R. Aarau: Sauerländer.
[30] Birnbaum, M. H. (2004). Human research and data collection via the Internet. Psychology, 55(1), 803.
[29] Bucholtz, M. (2003). Sociolinguistic nostalgia and the authentication of identity. Journal of Sociolinguistics, 7, 398–416.
[30] Eckert, P. (2003). Elephants in the room. Journal of Sociolinguistics, 7, 392–397.