

## CROWDSOURCING AND CITIZEN SCIENCE IN LINGUISTICS: NEW WAYS TO STUDY NON-STANDARD AND MINORITY LANGUAGES

BIRGIT ALBER, ANGELICA BONELLI, JOACHIM KOKKELMANS,

ANNA KATHARINA PILSBACHER

LIBERA UNIVERSITÀ DI BOLZANO/FREIE UNIVERSITÄT BOZEN

birgit.alber@unibz.it, angelica.bonelli2@student.unibz.it, jkokkelmans@unibz.it,

anna.pilsbacher@unibz.it

Received April 2025; Accepted July 2025; Published online December 2025

Crowdsourcing in linguistics has the potential of reaching a large number and a large variety of speakers, as well as creating a geographically fine-grained, well-distributed net of data points. These opportunities, as well as challenges of questionnaire design, data validity and data processing are discussed here in reference to the project AlpiLinK (Rabanus et al. 2025). The citizen science project VinKiamo Südtirol connected to AlpiLinK shows how involvement of schools in the crowdsourcing effort can help improve data quantity and quality and have an impact on the participants also in terms of the dissemination of scientifically grounded concepts of linguistic diversity.

*Keywords:* Crowdsourcing, Minority Language, Südtirol, Citizen Science, Crosslinguistic Studies

### 1. *Crowdsourcing and Citizen Science in Linguistics*

Online crowdsourcing has become increasingly popular in collecting linguistic data over the last decades. Even if we consider only projects focusing on the varieties of German crowdsourcing platforms or apps such as *AdA* (*Atlas der Alltagssprache*, Möller, Elspaß 2015), *gschmöis* (Hasse et al. 2021), *OeDA* (Vergeiner, Elspaß [forthcoming]) or *Verba Alpina* (Krefeld, Lücke 2016) have established themselves as important contributors to the documentation of language variation<sup>1</sup>. The focus of projects of this type has often been on variation in the lexicon or on dialectal variation as a consequence of sound change (but see the Marburg REDE project, Kasper, Pheiff 2023, for an example of a crowdsourcing project on morpho-syntactic variation in regional German). A further common characteristic of linguistic crowdsourcing projects is that responses by speakers are usually collected

<sup>1</sup> A recent addition is *DaBay*, a project crowdsourcing data on Bavarian dialects in Germany (<https://dialektapp.bayern/about>; last accessed July 1, 2025). See (Leemann 2021) for an overview of crowdsourcing projects documenting language variation in German; see (Gilles 2023) for results obtained with *Schn.ssen*, a smartphone app crowdsourcing Luxembourgish; see (Hilton, Leemann 2021) for an overview of linguistic crowdsourcing via smartphone apps in a variety of languages in Europe.

in written form and for varieties of a single language per project (but see *Verba Alpina*, Krefeld, Lücke 2016, which collects data for all languages of the Alpine region).

In this paper, the opportunities and challenges of linguistic crowdsourcing are discussed in reference to the project AlpiLinK (Rabanus et al. 2025; [alpilink.it](http://alpilink.it); last accessed October 31, 2025) and to one of its citizen science components, VinKiamo Südtirol (Siviero et al. 2025; [vinkiamo.projects.unibz.it](http://vinkiamo.projects.unibz.it); last accessed October 31, 2025). In the context of AlpiLinK, researchers aim to collect data for non-standard varieties and minority languages of Northern Italy. The Italian regions investigated by AlpiLinK include Piedmont, Aosta Valley, Lombardy, Veneto, Trentino-South Tyrol and Friuli-Venezia Giulia. One of the main goals of the project is to compare linguistic structures across the Romance, Germanic and Slavic varieties of the region to detect instances of contact-induced change. Besides the crosslinguistic perspective, a further innovative aspect of the project is that audio responses are collected for almost all elicitation tasks, obviating the problems involved in asking participants to write in their non-standard varieties. Furthermore, the elicited structures do not mainly concern the lexicon, but correspond to phonological, morphological and syntactic variables. Finally, the citizen science components of the project tread on unknown territory with their large-scale involvement of schools in the region in the scientific effort (Bertollo, Rabanus 2023).

Crowdsourcing in linguistics has the potential of reaching a large number and a large variety of speakers, as well as creating a geographically fine-grained, well-distributed net of data points (Kruijt et al. 2023). In comparison to traditional linguistic fieldwork, it is less costly and requires less time to collect a certain amount of data. This does not mean that crowdsourcing methodologies come without their own challenges. First and foremost, collecting data via crowdsourcing leads to a partial loss of control by the linguist. Once the questionnaire or data elicitation task is uploaded to the platform, participants fill it out on their own. No assistance from the linguistic expert and no clarifying questions, if tasks should be unclear, are possible. If an informant produces data which is interesting outside the limited scope of the questionnaires, the linguist will not be able to ask follow-up questions because informant anonymity removes the possibility of later contact. In this respect, crowdsourcing resembles the ‘indirect method’ by which Georg Wenker collected data on German dialects, sending letters to teachers asking them to translate forty sentences into the local variety with the help of their students in the 19<sup>th</sup> century. The response to this early crowdsourcing effort ultimately led to the creation of the maps of the *Sprachatlas des Deutschen Reichs* (Lameli 2014; Fleischer 2017). Wenker had to think carefully about how to pack all linguistic phenomena he was interested in into the forty sentences he proposed. In a similar fashion, modern crowdsourcing projects face the design of the questionnaire as a first challenge: it should simultaneously be short, so as not to bore or tire the informant (see e.g. Kruijt 2023 on feedback to the VinKo crowdsourcing project), and apt to elicit all the crucial contexts for the phenomenon the linguist is interested in. The indirect method used in linguistic crowdsourcing also bears similarities to the experimental methods used in psycholinguistics, although rather than eliciting responses to linguistic data, such as reaction times, the elicitation of the linguistic data itself is the main goal.

A second challenge faced by language documentation via crowdsourcing regards the validity of the collected data and the question whether it is comparable in quality to data collected in traditional fieldwork, where sessions with single informants may last several hours and involve continuous interactions between informant and fieldworker (see [Hilton, Leemann 2021] for an overview of the discussion on data validity in linguistic crowdsourcing). With respect to the varieties of Northern Italy, this issue has been addressed in detail by Kruijt et al. (2023) for data collected in the VinKo project, a precursor to the AlpiLinK project. Kruijt et al. (2023) compare data collected for three morpho-syntactic variables in a project making use of traditional fieldwork with data on the same phenomena collected in VinKo. The phenomena analysed regard the pronominal system of Tyrolean varieties, verb-postverbal subject agreement in the Romance varieties of Trentino and subject clitics in Veneto dialects. These phenomena have been investigated in both projects, the stimuli presented to the informants partly overlap and the regions for which data points exist are comparable, as is, to a large extent, the age profile of the informants. This means that the data which was elicited in the two projects can indeed be compared. Kruijt et al. (2023) observe similar patterns and a similar geographic distribution for data obtained through traditional fieldwork and data obtained via crowdsourcing. If anything, they note that crowdsourced data, given its sheer amount, can fill in gaps in the data set left by traditional fieldwork.

A challenge Wenker already had to contend with is the processing of the large amounts of data generated by crowdsourcing. Especially when audio data is collected at a large scale, manual transcription of the entire dataset is not feasible (so far, AlpiLinK has collected 57.109 audio files). Automatic speech recognition tools may help with the transcription of single words (see below) but are not yet able to yield transcriptions of longer utterances in non-standard varieties, since no significant transcribed corpora which could be used to train large language models such as Whisper (Radford et al. 2022) or Wav2Vec (Baevski et al. 2020) exist<sup>2</sup>.

A final challenge of crowdsourcing projects in the domain of linguistics concerns the recruitment of participants. People are interested in language, especially in the languages that they themselves speak. However, this general interest may not be enough for them to visit a website and fill out questionnaires, an activity which might take 30-60 minutes of their time. A reward system such as points obtained by filling out questionnaires are a possibility but might appeal only to some (especially younger) participants. While press campaigns increase the visibility of crowdsourcing efforts, they do not necessarily have a strong impact on the recruitment of participants. Conveying the goals of a project to the media can furthermore be challenging (see [Britain et al. 2018] for discussion). Turning the predicament into an opportunity, the AlpiLinK project has added the citizen science component VinKiamo to its scientific effort (Bertollo, Rabanus 2023; Siviero et al. 2025). One of the explicit goals of VinKiamo is to involve younger members of the language com-

---

<sup>2</sup> For an automated speech recognition model translating one of the AlpiLinK varieties, Tyrolean, into Standard German see AUGUSTA, developed at EURAC Research (Ducceschi, Franzini 2025). The ultimate goal, however, must be to generate orthographic representations of all non-standard varieties in AlpiLinK.

munities and train them to carry out interviews with older members of the community via the AlpiLinK platform.

In what follows, we discuss examples illustrating each of the points above. We discuss the complexity of designing crowdsourcing questionnaires, the validity of the data and the problems encountered in processing large amounts of audio data in section 2; the main focus of section 3 is the discussion of the opportunities and challenges encountered in the citizen science project Vinkiamo Südtirol.

## 2. Crowdsourcing – questionnaires, validity of data and processing of data

### 2.1 Questionnaire design

In AlpiLinK, linguists aim to elicit data for a variety of linguistic variables. Phonologists want to know details about the sound system, syntacticians aim to compare varieties, for instance, with respect to their pronoun and clitic system, and sociolinguists want to correlate various linguistic phenomena to sociolinguistic variables such as gender, age or location of the speakers (see [Blaxter, Britain 2021] on the relevance of metadata in crowdsourcing). One insight gained from collaborating with linguists from various backgrounds on the project at hand, is that questionnaires must be designed keeping the phenomenon which is to be studied in mind. No question or task fits all phenomena. For some phenomena, it is crucial that standard language text-stimuli be excluded, since any text in a standard language would influence speakers to reproduce the structures of the trigger text and avoid structures of their native variety. This is the case for phrasal verbs such as *taiàr zo*, literally ‘to cut down’ for the verb ‘to cut’, which are attested in the Romance varieties of Trentino and Ladin in similar contexts as in Germanic contact varieties (Bidese et al. 2016). A picture description task asking participants to describe pictures (as in Figure 1) ‘in one complete sentence’ allows to elicit sentences as in (1), containing a phrasal verb. It would have been difficult to elicit such structures if a translation task had been proposed instead. This is because a Standard Italian trigger sentence without phrasal verb as in (2) might not have resulted in a phrasal verb response, given the influence of the Standard Italian model (see [Mahowald et al. 2016] for a meta study on the process of syntactic priming in language production first described by Bock [1986]).

Figure 1 - Picture description task in AlpiLinK (task I01)



- (1) Participant U0382, from Cembra Lisignago (Trentino), age 17  
 el bekár le dre kel táia zo l salám  
 The butcher he=is behind that=he cuts down the salami  
 [The butcher is cutting the salami]
- (2) Standard Italian equivalent sentence  
 Il macellaio sta tagliando il salame  
 [The butcher is cutting the salami]

Besides the need to create phenomena-specific tasks, another lesson to be learnt from linguistic crowdsourcing is that tasks must be short, easy to understand and, possibly, fun to carry out (see [Kruijt 2021] for discussion). An example for the inclusion of fun in the AlpiLinK questionnaire is a word-creation task targeted at participants speaking Germanic varieties. In it, they are asked to form nouns from verbs using the circumfix *Ge- ...-e*, as in *schütteln* → *Ge-schüttl-e* [to shake → the shaking]. One of the goals of this part of the questionnaire is to test the productivity of this word-formation process. Therefore, participants are asked to form novel derived nouns, not yet attested in dictionaries. After reading an example task, participants are presented with a sentence containing the relevant verb (e.g. *Der Junge SCHÜTTELT die Bierflasche, bis das Bier herausläuft* [The boy shakes the beer bottle until the beer runs out]) and asked to fill the gap in the following sentence with the relevant derived noun (*Sein Freund sagt: Was ist denn das für ein ...?* [His friend says: What sort of ... is this?]). Because the circumfix *Ge- ... -e*, besides nominalising the verb, also expresses annoyance at the repetition of an action, participants can be heard giggling while performing the task. Thus, contrary to our fears, the task of creating new, not yet attested words, did not inhibit the participants from performing it, but rather added to the fun of leaving their data on the crowdsourcing platform.

## 2.2 Data validity

The validity of crowdsourcing data can be confirmed for the part of the AlpiLinK questionnaire dedicated to name truncation. In the literature, several ways of truncating names in Italian and German have been described (see [Alber 2010] for an overview of short names in Italian and [Arndt-Lappe 2018] for German). Italian for instance, allows a name such as *Francesca* to be shortened (among other possible structures) to the hypocoristics *Fra*, *France*, *Franci* or *Cesca*. Studies conducted using more traditional methodologies, i.e. mainly by testing patterns with participants recruited from the social networks of friends and family of the researchers, have shown that patterns of short names can be associated with the perceived age of name bearers. The conclusion, then, is that some short name patterns are older while others are of more recent formation. Alber and Kokkermans (2022) show that German speakers in South Tyrol perceive disyllabic short names preserving the stress of the base name (*Rési* for *Therésa*) as older than those preserving the left edge of the base name (*Mátthi* for *Matthias*). These, in turn, are perceived as older sounding than monosyllabic short names such as *Kath* for *Katharina*. A similar scale of diachronic change was established by Boschiroli (2017) for Italian in the Veneto region, with *Césca*

(for *Francésca*) exemplifying the oldest pattern, *Fránce* and *Fránci* more recent ones and *Fra* the most recent one. While the recruitment of participants from the domain of family and friends guarantees a certain amount of motivation to participate in the data collection, the numbers of participants recruited was rather low in both studies, reaching 25 German speakers in (Alber, Kokkermans 2022) and 13 Italian speakers in (Boschiroli 2017).

The addition of a task on name truncation to the AlpiLinK questionnaire, testing the perceived age of Italian and German short name patterns hypothesised in the previous studies, alleviated this recruiting problem. Results of all three data collection efforts can be compared because similar varieties (Germanic and Italian varieties in Northern Italy) were targeted and the average age of participants was similar as well. The task to elicit responses from participants had to be adapted to the crowdsourcing modality. In this more sophisticated task on AlpiLinK, participants had to choose the most appropriate short names for personae belonging to different age groups who interacted with each other in a dialogue (see [Alber et al. 2025] for details). The results confirm the diachronic scale proposed in the previous studies for both Italian and German: stress-preserving disyllabic short names (*Césca*, for *Francésca*) are older, disyllabic short names preserving the left edge of the base name (*Fránce*) are of more recent formation and monosyllabic short names (*Fra*) are the most recent ones. Crowdsourcing hence yields the same results as studies relying on more traditional methodologies. It does so, however, on the basis of an incomparably larger data set. Alber et al. (2025) can base their analysis of short names on the data of 163 speakers of a Germanic and 582 speakers of a Romance variety producing an overall number of 5513 items. In addition, the geographical distribution of the data is much more fine-grained, covering the whole Northern Italian region investigated in AlpiLinK. On the basis of this data, statistical analyses could be performed, and additional correlations, which previously had gone unnoticed, emerged. In Romance varieties, for example, *-i* suffixed short names such as *Fránci* for *Francésca* were correlated to the gender of the name bearers rather than their age: the probability that such patterns were used for female personae was much higher (more than four times as high) than the probability that they would be used for male personae (Alber et al. 2025, 24). Furthermore, a correlation between more recent patterns of the *Fra* type and their geographical distribution could be established for Romance short names. These patterns are more common in the Northwest of the AlpiLinK region. Their geographical origin might thus be hypothesised in the urban centres of Milan and Turin (Alber et al. 2025, 26). It is difficult to see how results of this type could have been obtained with more traditional types of fieldwork, at least assuming comparable human and financial resources invested in the research effort.

On the downside, given that crowdsourcing projects require short questionnaires which furthermore often test for various linguistic variables, name truncation patterns could only be tested for a restricted set of three Italian and three German base names in AlpiLinK, each proposed with a choice of three or four short names. This means that base names had to be controlled very carefully for frequency and average age of name bearers to avoid a disproportionate influence of the base name on the overall results. In sum, while linguistic crowdsourcing can yield large data sets as an output of tasks, in many cases the

input proposed to participants has to be more restricted than in the interviews commonly used in traditional fieldwork or in experimental settings for which participants are specifically recruited.

### 2.3 Data processing

While the name truncation questionnaire is designed as a multiple-choice test, for most tasks in AlpiLinK participants leave their data in the form of an audio file. This is necessary, since speakers of the Romance and Germanic varieties of the region are alphabetised in Standard Italian or German and are therefore not used to reading and writing their native non-standard variety. Audio responses make the tasks easier for participants and promise more authentic data to the researcher. The transcription or tagging of audio data is more complicated and less open to automatic processing than that of written texts, however. Here, it is important to take the variables of interest to the specific researcher into consideration. While it may be crucial to be able to analyse the acoustic properties of single sounds for a phonologist or phonetician, a syntactician may need a data set tagged for the presence or absence of phenomena such as clitic doubling. Thus, the question as to which data sets should be transcribed, in how much detail they should be described, and whether speech recognition systems are currently available for any of them, arises.

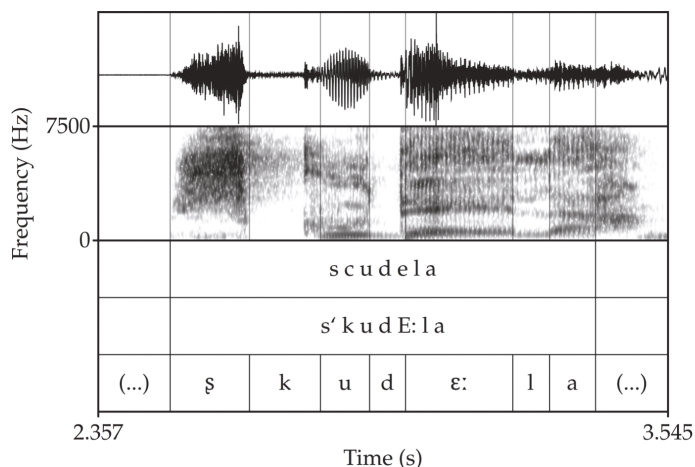
In the research team, automatic transcription of crowdsourced data started with data from the phonological questionnaire in the VinKo project, a precursor of the AlpiLinK project (Rabanus et al. 2025; Alber et al. 2018; Kruijt et al. 2023). The phonological questionnaire in VinKo consisted of a task where participants had to read words shown to them in an orthography created *ad hoc*<sup>3</sup> and accompanied by a translation into the related standard language (e.g. for Tyrolean the word *Städl* was proposed, together with its translation into Standard German *Heuschober* [hayloft or hayshed]). The reason that *ad hoc* orthographies were employed was to avoid that participants are primed by the form of the word in the standard language. However, participants noted in the feedback section at the end of the questionnaire that they were sometimes confused when the proposed word did not completely or partially correspond to the word of the participants' native dialect (Kruijt 2023). The goal of the task was to gain data from Romance and Germanic varieties to study the contrastive features involved in the obstruent system ([voice] vs. [spread glottis]), the realisation of /r/ (as alveolar or uvular) and the phonological processes in which sibilants are involved (voice assimilation or s-retraction). To this end, the task included words where obstruents, rhotics and sibilants occurred in all relevant contexts in the word. The questionnaire contained approximately 40 words, with some variation across varieties. In the context of the VinKo project 934 speakers responded to the phonological questionnaire and a total of 53.183 audio files were elicited.

---

<sup>3</sup> Specific orthographies had to be created for those varieties which do not have an orthographic norm. The research team took care that the phonemes of the variety corresponded to graphemes in a systematic way, but that at the same time the words were recognisable to speakers educated in a certain standard language.

In planning a semi-automated transcription of the audio files of VinKo or AlpiLinK, it is rather clear which responses to expect since the list of proposed words and sentences is known. This is an advantage with respect to spontaneous speech data. For the semi-automated transcription of the words of the phonological questionnaire in VinKo, forced alignment of segments with MAUS (Schiel 1999; 2015; Kisler et al. 2017) was chosen. The MAUS service takes an audio file together with a transcription of the expected segments in the IPA/X-SAMPA alphabet as an input and outputs a time-aligned transcription, i.e. a transcription containing the times (in centiseconds) at which each word and each segment starts and ends in the audio file. This type of forced alignment of acoustic data with segmental representation is particularly valuable for the acoustic analysis in phonetic and phonological studies. Once the data is time-aligned, it is easy to extract specific data, for instance all word-initial obstruents, and to analyse acoustic cues such as aspiration or voicing. A three-layered segmental representation was chosen where one layer consists in the sequence of segments as proposed to the speakers (the so-called ‘VinKoGraphy’), another layer contains the X-SAMPA-representation of the enunciation, which is easily converted into IPA (see example in Figure 2). A third layer, the ‘HistPhonGraphy’, contains additional information about segments in earlier diachronic phases allowing us to document sound changes occurring in some varieties. Figure 2 represents the oscillogram and the spectrogram (generated with Praat et al. 2025) of the Trentino dialect word *scudela* [bowl] with its VinKoGraphy on the top layer and X-SAMPA and time-aligned IPA transcriptions on the two bottom layers.

Figure 2 - *Oscillogram, spectrogram and time-aligned transcriptions of the Trentino dialect word scudela [bowl]*



While forced alignment via MAUS provides satisfactory results, it requires manual checking and, in some cases, correction of misaligned segments.

As of date, 2070 sound files from 5 different varieties have been force-aligned and manually corrected following this procedure. Once the whole corpus of available data has been aligned, the acoustic analysis of obstruents, rhotics and sibilants can start.

Researchers of the AlpiLinK team at the Free University of Bozen/Bolzano have worked to extend this procedure to the semi-automatic transcription of sentences. Since the AlpiLinK questionnaire provides Standard Italian or Standard German sentence prompts for tasks with written trigger sentences, a most probable output in the non-standard variety or minority language has been hypothesised and rendered in an *ad-hoc* orthography for each item. This proposed transcription is then presented to the researcher who can accept or modify it (e.g. when the word order in the audio differs from the hypothesised one) following a partially automatised workflow. Once a corpus of transcribed sentences is gathered, the sentences can be run through MAUS to obtain a time-aligned data set. Thus, while the transcription of audio data is not yet fully automatised, since manual correction is still required, most parts of the transcription process are. We hope that the creation of a corpus of transcribed single words and sentences can, in the future, be used as training data for large language models such as Whisper (Radford et al. 2022) or Wav2Vec (Baevski et al. 2020) making them, in turn, able to recognise larger chunks of speech in the non-standard or minority languages of the region. It also must be kept in mind that if transcriptions are generated, they should be consistent as well as acceptable to the speakers of each variety.

The experience of AlpiLinK shows that questionnaire design is a key issue in crowdsourcing: questionnaires must be reduced in size and tasks must be made easy and fun to carry out by the anonymous participant who might otherwise abandon the effort. Since only a small number of items can be presented, these must be selected with particular care. Tasks have to be designed anticipating the problems that participants might face, for instance by means of pilot runs. Finally, speech-to-text transfer can currently only be automatised for short utterances, in the case of non-standard or minority languages, since large language models such as Whisper or Wav2Vec are trained on standard varieties. These challenges come with some benefits, however. So far, all analyses testing crowdsourced data for its validity conclude that quality of data is not an issue. Our work shows that data collected with more traditional methodologies arrives at similar conclusions regarding the phenomena investigated compared to data collected through crowdsourcing. What's more, given the amount of data obtained through crowdsourcing it is easier to run statistical analyses on it. In some cases, this leads to previously undiscovered generalisations emerging, such as the correlation between the gender of a name and the form of the hypocoristic that is chosen for it.

### *3. Citizen Science and linguistic crowdsourcing – why and how*

Once a linguistic crowdsourcing platform has been set up, the recruitment of participants becomes a key concern. In order to fully exploit the advantages of the modality – i.e. the potential to gather a large amount of data for a set of geographically well-distributed data points – the number of participants in a crowdsourcing project should outnumber the infor-

mants of traditional linguistic fieldwork. Moreover, crowdsourcing of this type bears the risk that the data collected is seriously unbalanced with respect to the average age of participants: it will mostly be the younger generations who, as digital natives, are attracted to donating their data via an online platform. This was indeed the case for some of the data we collected for the precursor project VinKo. For the Germanic variety of Tyrolean, the average age of participants was 26 (as well as 90% female; Kruijt 2021; 2023). This particular unbalanced data set was the result of a recruitment campaign initiated among the university students attending a course taught by a team member, but the trend that certain subsets of a linguistic population are more likely to participate in online outreach projects has been observed previously (Hilton, Leemann 2021). As analyses should be based on large yet well-balanced data sets, new ways of reaching out to the communities of speakers become a necessity.

Besides the necessity to create a well-balanced corpus, involving speakers in the research effort creates the opportunity to disseminate the goals and results of linguistic research and to have an impact on the participants in terms of their attitudes to language in general and linguistic research in particular.

To pursue these goals, researchers at the University of Verona developed the citizen science project VinKiamo (Bertollo, Rabanus 2023), which was replicated with slight modifications in regional sub-projects such as VinKiamo Südtirol (Siviero et al. 2025), the citizen science project our research team developed at the Free University of Bozen/Bolzano.

### 3.1 VinKiamo Südtirol

The VinKiamo projects were born out of the realisation that press campaigns, although they enhance the visibility of the project, are not sufficient to recruit participants for all the varieties we are investigating with AlpiLinK. Though journalists were generally interested in the project and provided accurate and enthusiastic coverage, the press campaigns were not followed by a significant rise in collected data sets (for less positive experiences with press coverage in the context of *The English Dialects App* see [Britain et al. 2018]).

In developing VinKiamo Südtirol we intuitively followed the principles Tony McEnery (McEnery 2018) lists to achieve impact, which can be summarised as follows: (a) finding and keeping stakeholders; (b) aligning goals and workflows with stakeholders; (c) knowing you are being heard.

With respect to (a), secondary school students (age 14–19) and their teachers from the German and Ladin upper secondary schools of South Tyrol were contacted as stakeholders. Students were recruited as mediators and asked to help members of the older generation fill out the questionnaire on the AlpiLinK platform in order to obtain a more balanced corpus. We are convinced that teenagers, as digital natives, are able to bridge the digital divide with respect to the older generations and thus to help document the dialects of this age group.

While the AlpiLinK corpus is clearly benefiting from the VinKiamo projects in terms of quantity and quality of collected data, we think of students and teachers as true stakeholders (Price 2018), who can gain from participating in the research effort. In our conception of VinKiamo, one of the aims is to bring students in contact with goals and methodolo-

gies of linguistic research as well as with a concept of ‘language’ and ‘linguistic diversity’ free of the stereotypes often encountered in the communities. With respect to the latter point, the goal is to convey the idea that non-standard varieties and minority languages (together with many other varieties) contribute to the multilingual landscape of the region; that they can be the object of scientific studies because, similarly to the standard languages learnt at school, they have interesting structures; that it is not a problem, if children grow up with more than one language or dialect; that young speakers are not automatically ‘worse’ speakers, if their dialect differs from that of their elders, but that languages may change over time. In sum, the goal is to transmit a positive vision of the multilingual situation of the region allowing students and teachers to embrace it. The way VinKiamo is structured (see below), with students carrying out fieldwork and describing their experience in reports at the end of their participation, also gives them the possibility to develop organisational and team building skills as well as their ability to reflect on their experience.

With respect to (b), a pilot project with a secondary school (*Sozialwissenschaftliches Gymnasium*) in the city of Meran was initiated (Siviero et al. 2025), where interested teachers helped to understand how the project might fit into the tight school schedule. Together with the teachers, we decided which types of activities would be interesting for students and how the tasks that students performed could be evaluated for the school curriculum. Even though the goals of VinKiamo were developed by research teams at the universities of Verona and Bozen/Bolzano, they were adjusted in discussions with the students’ teachers who for instance contributed significantly to the proposals of how the students’ reports should be structured. Feedback questionnaires after the pilot helped us to understand shortcomings in this phase and to prepare for the main phase of the project.

With respect to (c), we thought of ways to measure the impact that the project might have had<sup>4</sup>. The number of questionnaires gathered on AlpiLinK following a VinKiamo Südtirol edition, or a change in the average age of participants, are obvious quantitative measures of the collaboration with the schools. However, these measures also list outcomes that benefit the researchers rather than the students in this collaboration. It is much more difficult to measure ‘whether you are being heard’. Did the students merely carry out fieldwork for AlpiLinK as an activity forced onto them by their curriculum without learning much about linguistic diversity or did they conclude the experience with a new view on language and linguistic diversity? Here, we present first results of attempts to understand this type of impact via a questionnaire we proposed to the students before and after each edition of the project and by reading the reports that students wrote about their experience.

### 3.1.1 Structure of VinKiamo Südtirol

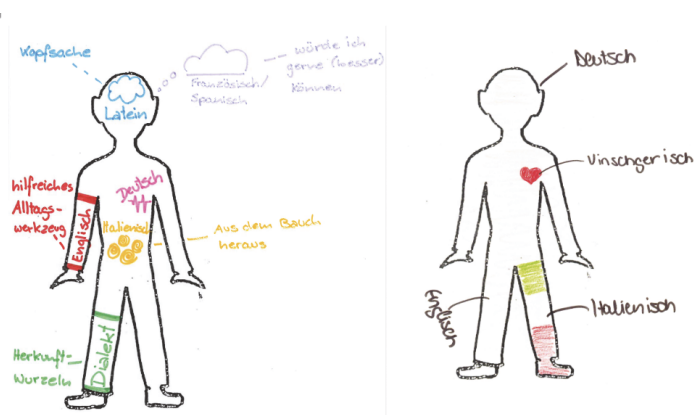
VinKiamo Südtirol editions have taken place four times, starting in the autumn of 2023, with the last edition taking place in March 2025 (once each school/university semester).

<sup>4</sup> The field of linguistics is less advanced in the development of methodologies to measure impact than fields such as environmental sciences, where various assessment methods of changes in the environmental knowledge, attitudes and behaviour as a consequence of citizen science projects exist (see [Somerville, Wehn 2022] for discussion; see [Wehn et al. 2021] for a proposal of general principles guiding impact assessment in citizen science projects).



After the presentation, students participated in a series of activities (organised in ‘stations’) which had the goal to illustrate different aspects of multilingualism. The activities were offered by volunteers from cultural institutes, by PhD and MA students of linguistics, or by linguists from the Free University of Bozen/Bolzano or other universities and institutions interested in the project. At each activity station, students had to perform a specific task. Representatives of the Mòcheno minority language, for instance, had prepared dictionaries and grammars of their language and asked students to put together a sentence in Mòcheno. At another activity station, called ‘Spot the 7 acoustic differences’, a researcher had put together spectrogram representations of cognates in Romance and Germanic varieties (‘international’ words like *garage* and *souvenir*). Students had to detect acoustic differences, e.g. aspiration on certain consonants, which are visible in the spectrograms and simultaneously audible in the corresponding audio fragments. At yet another activity station, ‘Does AI understand me?’, students were invited to test AI-driven language technologies such as Alexa or ChatGPT on their understanding and creation of texts in non-standard and minority languages. Figure 4 shows an example of the output of an activity where students represented their language repertoire in the form of a ‘language portrait’ (Krumm, Jenkins 2001).

Figure 4 - Language portraits of participants (VinKiamo Südtirol, October 2023)



One recurring type of language portrait (in Figure 4 to the right) sees non-standard varieties (*Vinschgerisch*) placed in the heart, while varieties deemed useful (Italian, English) occupy the legs. Standard German (*Deutsch*) apparently is connected to the head. More sophisticated representations such as the language portrait on the left place the non-standard variety in one leg, representing heritage and roots (*Herkunft, Wurzeln*), English, as a useful everyday tool (*hilfreiches Alltagswerkzeug*) in one arm, Italian in the guts (*Aus dem Bauch heraus* [from the belly]), German at the heart and Latin in the brain (*Kopfsache* [a matter of the brain]). French and Spanish are represented in a cloud outside of the body as ‘languages that I would like to know (better)’ (*würde ich gerne (besser) können*).

The activities were organised in the form of a game. Students had to collect stamps for each activity and were given a little gift as soon as they had collected a certain number of

stamps. In the feedback that students gave at the end of each VinKiamo edition, they particularly appreciated being at the University and interacting with university researchers, as well as the activity stations.

For the second phase of the project, members of the research team visited participating classes in their school environment. During these visits, questions such as how to access the AlpiLinK platform, how to recruit (older) informants, how to ensure the collection of quality data and the goals of the overall project were addressed. These visits gave students the possibility to get in touch with the principles of linguistic research such as preserving authenticity ('don't correct the informant') or the importance of a balanced data set.

During the last phase of the project, students set out in groups of two to engage in self-organised fieldwork. They recruited speakers of the older generation, helped them leave their audio data on the AlpiLinK platform and wrote a report about their experience. The report was supervised by their teachers who graded it, so that the experience could be recorded in the students' school curriculum. From the reports and the feedback of students it emerges that this was probably the most exciting phase of the whole project. Students saw their digital competences valued and appreciated, and at the same time discovered features of their own language and that of the speakers they interviewed (often their parents or grandparents) that they had not been aware of. As one of the students puts it in his/her report:

- (3) From students' reports (VinKiamo Südtirol, spring 2024)  
 Am interessantesten finde ich allerdings, dass mein Vater manchmal die Sätze anders gesprochen hat, als wie ich es tun würde. Auch wenn wir beide im gleichen Ort wohnen und aufgewachsen sind, kommt mir vor, dass ich dennoch ein wenig anders spreche.  
 [What I find most interesting is that my father said some sentences differently from what I would do. Even if we live and have grown up in the same place, I have the impression that I speak in a slightly different manner.]

They also mention that the interview often set off an intergenerational dialogue about language which otherwise might not have taken place.

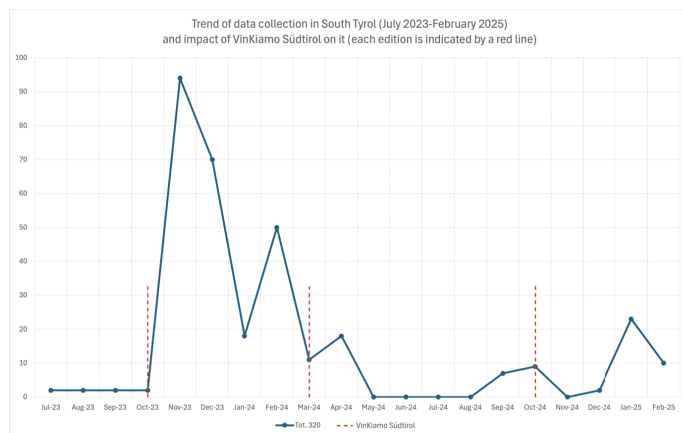
### 3.1.2 Determining impact of VinKiamo Südtirol

Quotes as that in (3) are signs that VinKiamo Südtirol has had an impact on the students that have participated in the project. But for the research team the issue of establishing and measuring impact is still open.

In terms of sheer numbers, we can observe that after each of the three VinKiamo Südtirol editions (the fourth has still to be concluded) the numbers of questionnaires have risen, as can be seen in the graph in Figure 5. Between the three months preceding the first VinKiamo Südtirol edition (August, September, October 2023) and the three months following it, the number of submitted AlpiLinK questionnaires from South Tyrol increased from six to 182. For March 2024, the increase is minor, since only one school class participated in this edition. For the October 2024 edition, the number of questionnaires in-

creased from 16 to 25 (+56.3%), comparing again a three-month period before and after the VinKiamo Südtirol edition.

Figure 5 - *Trend of number of questionnaires collected for AlpiLinK in South Tyrol after VinKiamo Südtirol editions*



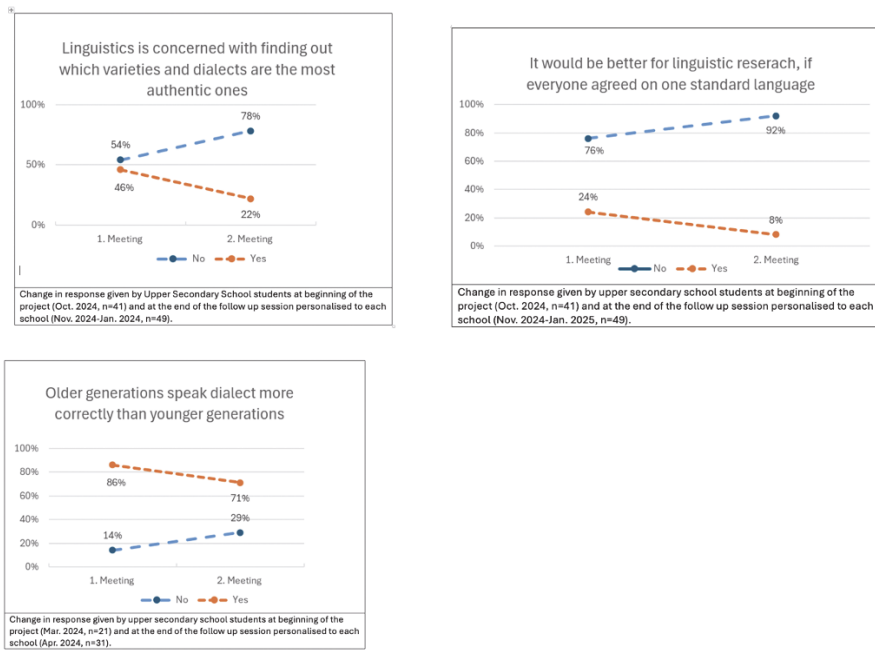
Not only has the number of questionnaires increased, but the quality of the language samples might be better than it would have been without VinKiamo Südtirol, as well. Although it is difficult to prove this, we deduce it from comments in the students' reports such as the following, which show the care that they put into carrying out the interviews:

- (4) From students' reports (VinKiamo Südtirol, autumn 2023)  
 Die Durchführung des gesamten Fragebogens dauerte etwa 60 Minuten, wobei wir zwischendurch kurze Pausen einlegten, um sicherzustellen, dass Frau \*\*\* konzentriert blieb.  
 [Answering the whole questionnaire took about 60 Minutes. During this time, we made short breaks to make sure that Ms. \*\*\* stayed concentrated]

Comments like these give also important feedback as to the length and difficulty of the AlpiLinK questionnaire.

In order to understand whether the students had 'heard us', they were asked to answer a short questionnaire both at the beginning and at the end of each edition of VinKiamo Südtirol. Among other, more fact-based questions about the linguistic landscape of the region, the questionnaire probed the students' attitude with respect to common stereotypes and misconceptions regarding non-standard or minority languages. Students were offered assurance that their answers were anonymous and would remain ungraded in a scholastic understanding of the term to ensure that they answered quickly and without the feeling of having to provide the 'correct' answer.

Figure 6 - Student questionnaire documenting change in responses to stereotypes/misconceptions concerning non-standard languages



First results show that these attitudes have indeed changed between the beginning and the end of each edition. As the diagrams in Figure 6 show, the responses to statements such as ‘Linguistics is concerned with finding out which varieties and dialects are the most authentic ones’, or ‘It would be better for linguistic research, if everyone agreed on one standard language’, as well as ‘Older generations speak dialect more correctly than younger generations’ have clearly changed after participating in VinKiamo Südtirol.

A third way to measure the impact of VinKiamo Südtirol is to analyse the reports that students wrote on their experience searching for statements that reveal an increase in their understanding of the goals and methodologies of linguistic research as well as with respect to the features of the non-standard and minority languages of the region. This has not yet been done in a systematic fashion since the reporting phase for the third and fourth edition is not finished but extracts from students’ reports during the first and second edition of VinKiamo Südtirol show that this is indeed the case. The reports contain general observations of the multilingualism of certain communities, as in (5a), but also considerations about how to interact with informants without endangering the authenticity of the data, as in (5b). In certain cases, students were quite explicit about what they learnt during the project (5c).

- (5) Extracts from students' reports (VinKiamo Südtirol autumn 2023 and spring 2024)

(a) Dër interessant él da fá n'intervista te n pice paisc, sciöche la Ila olache al vëgn baié plü lingac: talian, todësch y ladin y oramai ince inglesc. An pó osservé y se intëne che nia dötes les porsones ne é bones da baié n ladin nët.

[It is very interesting to make an interview in a small village such as La Ila, where several languages are spoken: Italian, German and Ladin and nowadays also English. It can be observed and understood that not every person is able to speak correct Ladin.]

(b) In bestimmten Sätzen fiel mir auf, dass mein Vater auch denken müsste wie [er] dieses Formulieren soll um den eigenen Dialekt zu verwenden. Beeinflusst habe ich mein Sprecher nicht, denn ich habe ihn in Ruhe diesen Fragebogen durchführen gelassen.

[In certain sentences I noticed that my father also had to think how he should phrase this in order to use his own dialect. However, I have not influenced my speaker, because I let him answer this questionnaire in peace]

(c) Um ehrlich zu sein habe ich mein ganzes Leben lang den Dialekt als etwas 'Schlechtes' wahrgenommen, als etwas, das die deutsche Hochsprache degradiert. [...] Als wir aber Ende Oktober eine Einführung zu Mehrsprachigkeit und Dialekten an der Universität Brixen erhalten haben, wurden meine Überzeugungen kurzerhand auf den Kopf gestellt. Uns wurde mitgeteilt, dass ein Dialekt sehr wohl ein vollwertiges Sprachsystem mit einem eigenen Lautsystem, einer Wort- und Satzstruktur und Wortschatz ist. Auch wenn das Regelwerk nicht so rigide wie in der Hochsprache ist und Dialekte eher in einem informellen Kontext Verwendung finden, sind sie keineswegs als 'falsch' anzusehen.

[To tell the truth, for my whole life I have perceived the [Tyrolean] dialect as something 'bad', as something that degrades the German standard language. [...] But when we received at the end of October an introduction to multilingualism and dialects at the University in Brixen, my beliefs were quickly turned upside down. We were told that a dialect is a full-fledged linguistic system with its own sound system, a word and sentence structure and a vocabulary. Even if the system of rules is not as rigid as in the standard language and dialects are used rather in informal contexts, they cannot be considered 'wrong'.]

The student who wrote the (nine-page) report from which (5c) is extracted proceeds in his/her text with an in-depth analysis of the linguistic features distinguishing their own Tyrolean dialect from Meran from their mother's (Tyrolean dialect from Vinschgau) and their father's (Tyrolean dialect from Passeiertal), as they emerged in the interview. The analysis refers to differences in the sound systems (alveolar vs. uvular /r/, apocope of final schwa) and in the morphology (past participle formation). It is impressive that exposure to a project like VinKiamo Südtirol could trigger observations which are not too distant from those that a trained linguist would propose.

To summarise, the VinKiamo project shows that the combination of linguistic crowdsourcing with citizen science projects has multiple advantages. Researchers obtain data sets that are richer, both quantitatively and qualitatively speaking. At the same time, they have the opportunity to disseminate goals, methodologies and results of their research and, in certain cases, to have real impact on communities outside of the academic environment. The interaction of linguists with members of the public such as schools, furthermore increases the visibility of linguistic diversity in general and of communities of endangered languages in particular. We recognise that the question of how to measure the actual impact of projects of this type is an aspect that requires more attention. At the moment it seems to us that a combination of quantitative measures, such as an increase of the corpus after citizen science activities, together with a change in responses to stereotypical opinions about non-standard and minority languages as well as a content analysis of participants' reports might be the best approach (see [Wehn et al. 2021] for a similar proposal).

#### 4. *Conclusions*

In this paper, the projects AlpiLinK and VinKiamo Südtirol have been presented as examples of how crowdsourcing and citizen science projects can be implemented in the domain of linguistics. With respect to AlpiLinK, we have discussed the necessity to insert tasks into the online questionnaires which are both short, easily executed and interesting to participants, as well as tailored to elicit specific phenomena. The validity of crowdsourced data is a concern, since researchers do not have direct access to informants. In our work, however, crowdsourced data yielded descriptions of linguistic phenomena comparable to those elicited with traditional methodologies. What is more, given the larger data sets generated through crowdsourcing, generalisations on linguistic patterns might emerge which traditional fieldwork is not able to uncover, at least not with a comparable number of researchers and similar financial support. The processing of large data sets remains an open issue, especially if the data takes the form of audio files. Tools such as MAUS are useful and large language models might solve some problems, as soon as they are able to also recognise non-standard or, more generally, low-resource languages. With respect to projects in the domain of citizen science, we have reported on the experiences during the four editions of VinKiamo Südtirol. In the interaction of the research team with German and Ladin schools in South Tyrol both partners benefitted from the cooperation. The AlpiLinK corpus increased both in quantity and in quality as a direct consequence of the citizen science projects around VinKiamo while students had the opportunity to get to know linguistic research up close. An open issue concerns the question of the impact that citizen science projects in the domain of linguistics can leave on the participants. We propose to approach this issue from different angles, documenting the results in terms of data increase, as well as monitoring the change of attitudes towards language expressed by participants during the project.

### *Acknowledgement of support*

This research has been carried out within the PNRR research activities of the consortium iNEST (Interconnected North-East Innovation Ecosystem) funded by the European Union Next-GenerationEU (Piano Nazionale di Ripresa e Resilienza (PNRR) Missione 4 Componente 2, Investimento 1.5 D.D. 1058 23/06/2022, ECS\_00000043).

This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

### *References*

- Alber, Birgit. 2010. "An Exploration of Truncation in Italian." In *Rutgers Working Papers in Linguistics* 3, edited by Peter Staroverov, Daniel Altshuler, Aaron Braver, Carlos A. Fasola, Sarah Murray, 1–30. Rutgers Linguistics. <https://doi.org/10.7282/T33B5XT9>.
- Alber, Birgit, Sabine Arndt-Lappe, Joachim Kokkelmans. 2025. "The Predictability of Name Truncation: Factoring in Language Change." *Catalan Journal of Linguistics* 24 (1): 7–39. <https://doi.org/10.5565/rev/catjl.467>.
- Alber, Birgit, Ermenegildo Bidese, Jan Casalicchio, Patrizia Cordin, Antonio Mattei, Andrea Padovan, Stefan Rabanus, Alessandra Tomaselli. 2018. "VinKo, Versione 2." In *Lo Spazio Comunicativo Dell'Italia e Delle Varietà Italiane (Korpus Im Text 7), Versione 91*, edited by Roland Bauer, Thomas Krefeld. <https://www.kit.gwi.uni-muenchen.de/?p=13739&v=2> (last accessed October 31, 2025).
- Alber, Birgit, Joachim Kokkelmans. 2022. "Typology and Language Change: The Case of Truncation." *Isogloss. Open Journal of Romance Linguistics* 8 (2): 1–17. <https://doi.org/10.5565/rev/isogloss.124>.
- Arndt-Lappe, Sabine. 2018. "Expanding the Lexicon by Truncation: Variability, Recoverability, and Productivity." In *Expanding the Lexicon*, edited by Sabine Arndt-Lappe, Angelika Braun, Claudine Moulin, Esme Winter-Froemel, 141–70. De Gruyter. <https://doi.org/10.1515/9783110501933-143>.
- Baevski, Alexei, Henry Zhou, Abdelrahman Mohamed, Michael Auli. 2020. "Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations." arXiv. <https://doi.org/10.48550/arXiv.2006.11477>.
- Bertollo, Sabrina, Stefan Rabanus. 2023. "VinKiamo: ein Citizen-Science-Projekt für Schulen zur Förderung von (sprach-) übergreifenden Kompetenzen." *Alsic* 26 (1). <https://doi.org/10.4000/alsic.7076>.
- Bidese, Ermenegildo, Jan Casalicchio, Patrizia Cordin. 2016. "Il ruolo del contatto tra varietà tedesche e romanze nella costruzione «verbo più locativo»." *Vox Romanica* 75: 116–42.
- Blaxter, Tamsin, David Britain. 2021. "Hands off the Metadata!: Comparing the Use of Explicit and Background Metadata in Crowdsourced Dialectology." *Linguistics Vanguard* 7 (s1): 20190029. <https://doi.org/10.1515/lingvan-2019-0029>.
- Bock, J. Kathryn. 1986. "Syntactic Persistence in Language Production." *Cognitive Psychology* 18 (3): 355–87. [https://doi.org/10.1016/0010-0285\(86\)90004-6](https://doi.org/10.1016/0010-0285(86)90004-6).
- Boersma, Paul, David Weenink. 2025. "Praat: Doing Phonetics by Computer." <http://www.praat.org/> (last accessed October 31, 2025).

- Boschioli, Laura. 2017. "Bestimmung des Alters italienischer Trunkierungsmuster." Term paper, University of Verona.
- Britain, David, Marie-José Kolly, Adrian Leeman. 2018. "Using Impact to Make Impact? Experiences from a Dialect Crowdsourcing Project." In *Applying Linguistics: Language and the Impact Agenda*, edited by Dan McIntyre, Hazel Price, 83–111. Abingdon: Routledge.
- Ducceschi, Luca, Greta Franzini. 2025. "Augusta: ASR model for South Tyrolean dialect transcription into Standard German." Zenodo. <https://doi.org/10.5281/ZENODO.15553914>.
- Fleischer, Jürg. 2017. *Geschichte, Anlage und Durchführung der Fragebogen-Erhebungen von Georg Wenkers 40 Sätzen: Dokumentation, Entdeckungen und Neubewertungen*. Deutsche Dialektgeographie 123. Hildesheim: Georg Olms Verlag.
- Gilles, Peter. 2023. "Regional variation, internal change and language contact in Luxembourgish: results from an app-based language survey1." *Taal en Tongval* 75 (1): 29–57. <https://doi.org/10.5117/TET2023.1.003.GILL>.
- Hasse, Anja, Sandro Bachmann, Elvira Glaser. 2021. "Gschmöis – Crowdsourcing Grammatical Data of Swiss German," January. <https://doi.org/10.5167/UZH-203209>.
- Hilton, Nanna Haug, Adrian Leemann. 2021. "Editorial: Using Smartphones to Collect Linguistic Data." *Linguistics Vanguard* 7 (s1): 20200132. <https://doi.org/10.1515/lingvan-2020-0132>.
- Kasper, Simon, Jeffrey Pheiff. 2023. "From Dialect Syntax to Regional Language Syntax. Syntactic Variation between Dialect and Standard." *Zeitschrift für Dialektologie und Linguistik* 90 (1): 64–96. <https://doi.org/10.25162/zdl-2023-0003>.
- Kisler, Thomas, Uwe Reichel, Florian Schiel. 2017. "Multilingual Processing of Speech via Web Services." *Computer Speech & Language* 45: 326–347. <https://doi.org/10.1016/j.csl.2017.01.005>.
- Krefeld, Thomas, Stephan Lücke. 2016. "Verba Alpina." Ludwig-Maximilians-Universität München. <https://doi.org/10.5282/VERBA-ALPINA>.
- Kruijt, Anne. 2021. "Crowdsourcing Language Contact. Pronoun and Article Morphology in Trentino-South Tyrol and Veneto." PhD diss., University of Verona.
- Kruijt, Anne. 2023. "VinKo. Final Report." Technical Report. Verona: University of Verona.
- Kruijt, Anne, Patrizia Cordin, Stefan Rabanus. 2023. "On the Validity of Crowdsourced Data." In *Corpus Dialectology*, edited by Elissa Pustka, Carmen Quijada Van Den Berghe, Verena Weiland, 10–33. Amsterdam: John Benjamins Publishing Company. <https://doi.org/10.1075/scl.110.01kru>.
- Kruijt, Anne, Stefan Rabanus, Marta Tagliani. 2023. "The VinKo Corpus. Oral Data from Romance and Germanic Local Varieties of Northern Italy." In *Neue Entwicklungen in Der Korpuslandschaft Der Germanistik. Beiträge Zur IDS-Methodenmesse 2022*, edited by Marc Kupietz, Thomas Schmidt, 1st ed., 203–212. <https://doi.org/10.24053/9783823396024>.
- Krumm, Hans-Jürgen, Eva-Maria Jenkins, eds. 2001. *Kinder und ihre Sprachen – lebendige Mehrsprachigkeit: Sprachenporträts*. Wien: Eviva.
- Lameli, Alfred. 2014. "Georg Wenker auf dem Weg zum Sprachatlas des Deutschen Reichs." In *Schriften zum Sprachatlas des Deutschen Reichs. Band 3: Erläuterungen und Erhebungsmittel zu Georg Wenkers Schriften*, edited by Alfred Lameli, 1–69. Deutsche Dialektgeographie, Band 111 3. Hildesheim: Georg Olms Verlag.
- Leemann, Adrian. 2021. "Apps for Capturing Language Variation and Change in German-Speaking Europe: Opportunities, Challenges, Findings, and Future Directions." *Linguistics Vanguard* 7 (s1): 20190022. <https://doi.org/10.1515/lingvan-2019-0022>.

- Mahowald, Kyle, Ariel James, Richard Futrell, Edward Gibson. 2016. "A Meta-Analysis of Syntactic Priming in Language Production." *Journal of Memory and Language* 91 (December): 5–27. <https://doi.org/10.1016/j.jml.2016.03.009>.
- McEnery, Tony. 2018. "Reflections on Impact." In *Applying Linguistics: Language and the Impact Agenda*, edited by Dan McIntyre, Hazel Price, 29–40. Abingdon, Oxon/New York: Routledge.
- Möller, Robert, Stephan Elspaß. 2015. "21. Atlas zur deutschen Alltagssprache (AdA)." In *Regionale Variation des Deutschen*, edited by Roland Kehrein, Alfred Lameli, Stefan Rabanus, 519–540. De Gruyter. <https://doi.org/10.1515/9783110363449-022>.
- Price, Hazel. 2018. "Navigating the Peripheries of Impact. Public Engagement and the Problem of Kneejerk Linguistics." In *Applying Linguistics: Language and the Impact Agenda*, edited by Dan McIntyre and Hazel Price, 41–52. Abingdon, Oxon/New York: Routledge.
- Rabanus, Stefan, Anne Kruijt, Birgit Alber, Ermenegildo Bidese, Livio Gaeta, Gianmario Raimondi. 2025. "AlpiLinK Corpus 1.2.0." Zenodo. <https://doi.org/10.5281/ZENODO.15129710>.
- Radford, Alec, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, Ilya Sutskever. 2022. "Robust Speech Recognition via Large-Scale Weak Supervision." arXiv. <https://doi.org/10.48550/arXiv.2212.04356>.
- Schiel, Florian. 1999. "Automatic Phonetic Transcription of Non-Prompted Speech." In *Proceedings of the XIVth International Congress of Phonetic Sciences: ICPHS 99*; San Francisco, 1–7 August 1999, edited by John J. Ohala, 607–610. San Francisco. <https://doi.org/10.5282/UBM/EPUB.13682>.
- Schiel, Florian. 2015. "A Statistical Model for Predicting Pronunciation." In *International Congress of Phonetic Sciences*. <https://api.semanticscholar.org/CorpusID:30405308>.
- Siviero, Emily, Birgit Alber, Joachim Kokkelmans. 2025. "Dialektforschung und Schule: Das Projekt VinKiamo Südtirol." In *Dialekt in der Lehre: Sprachdidaktische und varietätenlinguistische Perspektiven*, edited by Christiane Hochstadt, Anny Schweigkofler Kuhn, 1. Auflage, 165–180. Tübingen: Narr Francke Attempto. <https://doi.org/10.24053/9783381109029>.
- Somerwill, Luke, Uta Wehn. 2022. "How to Measure the Impact of Citizen Science on Environmental Attitudes, Behaviour and Knowledge? A Review of State-of-the-Art Approaches." *Environmental Sciences Europe* 34 (1): 18. <https://doi.org/10.1186/s12302-022-00596-1>.
- Vergeiner, Philip, Stephan Elspaß. In press. "OeDA! Vom Nutzen Einer Sprach-App Für Die Erforschung Österreichischer Dialekte." In *Festschrift Für Peter Ernst*, edited by Marietta Calderón, Stephan Gaisbauer, Sandra Herling. Meßkirch: Gmeiner.
- Wehn, Uta, Mohammad Gharesifard, Luigi Ceccaroni, Hannah Joyce, Raquel Ajates, Sasha Woods, Ane Bilbao, Stephen Parkinson, Margaret Gold, Jonathan Wheatland. 2021. "Impact Assessment of Citizen Science: State of the Art and Guiding Principles for a Consolidated Approach." *Sustainability Science* 16 (5): 1683–1699. <https://doi.org/10.1007/s11625-021-00959-2>.

