Chapter 7

Validity of crowd-sourced minority language data: Observing variation patterns in the Stimmen recordings

Nanna Hilton[a]

[a]University of Groningen

Minority languages are underrepresented in linguistic research, and a possible reason for this is the lack of accessible speech recordings from lesser-used languages. This paper considers the usability of crowd sourced minority language data for research, focussing on the speech recordings and reported dialect knowledge collected with the smartphone application Stimmen, ('Voices' in Frisian). In this paper variation patterns in Frisian speech data from the Stimmen project (2017-2019) are compared with findings from previous sociolinguistic research in Fryslân. The comparison focusses on three phonological variables in Frisian speech: the coda cluster /sk/; the vowel in 'eye', and the realisation of post-vocalic coda /r/. The analysis conducted on the crowd sourced recordings show the same variation patterns as that of previous research, giving validity to the data for use in sociolinguistic studies of variation and change in minority language settings. This gives hope for future research that relies on the collection of speech samples in a remote capacity.

1 Background

1.1 The European bias in (socio)linguistics

The 'Stimmen fan Fryslân' project, Stimmen for short, is a citizen science project with the aim of collecting speech data from lesser used languages for research and technological development. While we estimate that there are 5,000-7,000 languages used across the globe, publications within the field of linguistics largely

represent (speakers of) large languages, and especially English. Nagy & Meyer-hoff (2008) conclude, for example, that in 449 publications and presentations in four key outlets for research in (variationist) sociolinguistics 50-70% concerns the English language.

For a field such as sociolinguistics, that deals predominantly with diversity in language, how and why language changes, and what the linguistic and social consequences of such changes are (cf. Weinreich et al. 1968), it is clear that biases in the empirical foundations are problematic. This is a likely reason why in recent decades more studies have appeared that incorporate the social reality outside that of majority populations into sociolinguistic theory. Stanford & Preston (2009) and Stanford (2016) point out how minority language communities offer a welcome chance for variationist sociolinguistics to revisit principles of linguistic variation and change, most of which take a European, or Anglo-American, linguistic and societal constellations for granted. Nagy (2009: 400) notes in her study of Faetar that minority languages without a codified and accepted standard give sociolinguists an opportunity to do research without influence from a standard language ideology. Studies outside the Anglo-American sphere also reconsider the arrangements of the traditional social variables class, gender and age: Noglo (2009), for example, shows how factors such as monetary income, property ownership and fortune are not part of what constitutes social class in many societies, providing a case study where ethnicity and community membership is used to define layers of social hierarchies. Brunelle (2009), similarly, argues how the Labovian principles of gender and language cannot be assumed before carefully considering the type of access to political and financial privilege and prestige that women have in specific communities outside the US and UK. O'Shannessy (2009) indicates in her study of language change in northern Australian indigenous languages that the variationist sociolinguistic principles of age and apparent time cannot be assumed in communities where contact with a majority language is widespread and the younger generations have a higher degree of bilingualism.

Perhaps more than for other linguistic disciplines the study of social variation in language is reliant on previous descriptive work having been done for the language in question. A researcher wanting to enter a community to understand the social role of language is reliant on already being a speaker of the language, or on available teaching material for acquiring the language. Arguably, linguistic documentation is also key for a number of applications of our linguistic research: the development of language technology, monitoring of learning abilities, and forensic analyzes concerning language, cannot exist without proper documentation. It is problematic then, that the majority of the world's languages are still

under-documented (Hammarström and Nordhoff, 2011). More than a thousand varieties are only described with a wordlist, a text collection, or not at all. One of the challenges the linguistic community is faced with as of now is that available data, and particularly speech data, from lesser used languages is hard to come by. While collections of speech can often be found on the internet, it may not be in a format that lends itself well for research, without annotation and translations, for example, or of poor quality due to recordings made in noisy environments.

1.2 Citizen science and crowd sourcing language data

A way to increase representation of under-resourced language communities in science is to engage in "citizen science": the participation of the general public alongside scientific researchers, in some or all steps of the research project (cf. Bonney et al. 2016). While the label "citizen science" is relatively recent, the practice itself is not. In linguistics, the contribution of the general public as providers of empirical data has a history that goes back centuries. Early descriptive linguistic work, such as dialectological studies where speakers of localised varieties provided data for the analysis of regional variation in language (e.g. Wenker 1881), can be seen as examples of "citizen science". However, recent technological developments have made public involvement in linguistic research substantially easier to facilitate, see, for instance, examples of dialectological surveys being conducted using the internet such as Vaux (2004) and Möller & Elspaß (2015). Smartphones offer further functionalities that can be employed to collect, store, share and analyze large amounts of language. Applications for iOS or Android have been used to, e.g. collect speech for development of language technology (De Vries et al. 2014), to make high-quality recordings for acoustic phonetic research (De Decker & Nycz 2011), or to collect reported language use for the creation of new dialect maps (Kolly & Leemann 2015). Yet, an issue with remote crowd sourcing of speech data remains its anonymous nature and the lack of control on the side of the researcher.

1.3 What constitutes "good" crowd sourced minority language data?

A commonly held conception about crowd sourced data is that it is problematic for research purposes due to the lack of control that the researcher can execute over the data collection process and context. Researchers could receive contributions that are done merely in jest, there could be participants who contribute several times, there could be contributors who have not read instructions, or contributions that do not meet the expectations of the researcher in another way.

Studies of the validity and reliability of crowd sourced data indicate that these worries are not entirely justified. Lind et al. (2017) find in their study of "crowd-coding" (coding of content analysis of news articles in German) that paid volunteers (through the platform Mturk) performed comparably to five research assistants, but that there was variability across different types of tasks, and within the group of coders. Horn (2019) concludes in a study of whether non-expert coders can perform quite complex content analysis of political messages, and conclude that crowd sourced coding is equally reliable to that of experts. However, there seems to be a limit to task complexity before the validity and reliability suffers: Shing et al. (2018) compare experts' and crowds' (CrowdFlower) annotations of online postings for authors' suicide risk, and find, unsurprisingly, that the experts outperform the crowd, even when the crowd has been given detailed task instructions. One previous study has considered the validity of crowd sourced linguistic judgements for the purpose of studying language variation and change. Leemann et al. (2016) find that the dialect judgements crowd sourced with the smartphone based Dialekt Äpp correlate to a high degree with speech samples collected with traditional dialectological methods.

The findings in Leemann et al. (2016) bode well for a study of the validity of crowd sourced speech data using smartphones. However, a number of validity-related concerns exist specifically for the crowd sourcing of minority language data. Shameem (1998) points out, for example, that self-reported minority language proficiencies do not always compare to actual test results. In her study of Indo-Fijians living in New Zealand she finds some over-reporting of oral minority language proficiency. Contrarily, Nicholas (1988) finds that minority children typically under-report their knowledge of the heritage language in surveys of British language diversity. There is also reason to question whether surveys designed for the whole of a minority speech community does in fact reach all those that could interest the researcher. "New speakers", for instance, could experience a lack of ownership and legitimacy in the use of the minority language (cf. O'Rourke & Ramallo 2013) and not see themselves as "real" minority language speakers.

Minority language communities cannot be approached with the traditional dialectological methods employed for instance in Leemann et al. (2016) either. While many minority languages may have a written standard, the acceptance levels of the standard language can be highly variable within the communities. Dorian (1978: 592) concludes that a majority language, in her case English, can act as the formal register in bilingual speakers who are less proficient in a minority variety, Gaelic in the case at hand. In communities where the minority language is not used throughout the educational system the majority language

generally enjoys higher instrumental value. This means that the reliance of written language in crowd sourcing of minority language can be problematic, with respondents wanting to perform well and give the most correct form as a response. This could lead to questionable outcomes, complicating the design of the study.

This study is the first to consider the validity of crowd-sourced speech data from a minority language. It does so by comparing the speech that users have themselves, as the citizen scientists, categorised as "Frisian", collected in a picture naming task in the Stimmen application, with previous data collected in Fryslân to investigate language variation and change. Additionally, data from a gamified dialect task for Frisian is used for the comparison. Before moving on to the analysis of the speech data the design and the context of the study is laid out below.

2 The current study

2.1 Frisian and the basis for data comparison

Fryslân is the only officially multilingual province in the Netherlands, and is a region in which several (regional and migrant) minority languages are used. Frisian-Dutch bilingualism is widespread in the province: 75% of the 9915 inhabitants that the Province of Fryslân surveyed in 2015 report being able to speak Frisian (TaalAtlas 2015: 3). Three quarters of the population would correspond to some 485,000 speakers of Frisian in Fryslân (Centraal Bureau voor de Statistiek 2018). These speakers are all presumed bilingual as secondary schooling is taught partially, or only, in Dutch. The province Fryslân is also home to mixed languages (results of long-term contact between Dutch and Frisian) 'Bildts' and 'Town Frisian' (van Bree 1994, Hoekstra & van Koppen 2000), as well as to varieties of another West-Germanic language family: Low Saxon, spoken along the southern border of the province of Fryslân.

An assumption for sociolinguistic work on Frisian is that the language is converging towards Dutch on all linguistic levels, cf. Breuker (1992) and De Haan (1997). The study of contact phenomena between Frisian and Dutch has been given considerable attention in Frisian linguistics, e.g. Sjölin (1976), yet generally in an introspective manner, or with anecdotal evidence. Very few empirical studies of speech variation in Frisian exist, and those that have been conducted have all been studies of phonetics and phonology. They indicate no evidence of convergence between Frisian and Dutch: Feitsma et al. (1987, data collected

1982–1984) is the oldest study available, looking at sandhi phenomena in an apparent time study. The study finds no signs of convergence between Frisian and Dutch, but rather indications of divergence for some phonetic variables. Van Bezooijen (2009) investigates the variation and change in pronunciation of /r/ in Frisian, which traditionally has the alveolar trill, and concludes that the approximant variant of /r/ that is gaining ground in the Dutch speech community does not occur at all in the Frisian data (Van Bezooijen 2009: 312). Finally, Nota et al. (2016) find gender and age differences in the realisation of intonation contours in 40 bilingual Frisian-Dutch speakers, but no direct indications that Dutch and Frisian intonation contours are converging.

In this paper, the research question does not directly concern the convergence between Frisian and Dutch, but rather asks what the usability, that is, the validity, of Frisian data collected by crowd sourcing is for conducting sociolinguistic analyzes. However, the data will be compared to other work concerned with the question of whether Frisian is changing towards Dutch and can in that sense possibly corroborate some previous findings. One of the data sets used for comparison is that of Van Bezooijen (2009), used to compare variation patterns in the pronunciation of (r). Furthermore, variation in the consonant coda cluster (sk) is looked at alongside data from Hilton & Weening (2014). Variation in the vowel in the word *each* "eye" is also considered, by comparing crowd-sourced data to that from an ongoing study by Stefan, Klinkenberg & Versloot (cf. Stefan et al. 2014).

2.2 The citizen science project "Stimmen"

The crowd sourced data discussed in this paper was collected with the smartphone application Stimmen. The app was part of the larger "Stimmen fan Fryslân" project funded by the program "Lân fan taal" in the European Capital of Culture project for Leeuwarden 2018. The citizen science components of the Stimmen project consisted of collaboration with twelve secondary schools in Fryslân as well as numerous online and offline activities held throughout the year of 2018. An estimated 2,000 respondents were reached through the offline activities, that gave the researchers an opportunity to share findings with the general public, and, crucially, for the public to approach the researchers to post questions about language. Since the start of the project the smartphone application has been downloaded 6,039 times (iOS: 2,989; Android: 3,050. In 2017, 28% of smartphones in the Netherlands was an iPhone (Steemers et al. 2017), but the application could also be downloaded on an iPad which may eschew the iOS user numbers), yet this is not equivalent to the amount of people who have submitted data through

the application. The dialect quiz has been used a total of 15,131 times. The discrepancy between the two figures can be explained by the fact that the dialect quiz, see details below, is also available as an easily shared web application on stimmen.nl.

3   The Stimmen smartphone application

The Stimmen application is available for Android or iOS and consists of the following components: a start screen with choice of interface language (English, Dutch, or Frisian), a tutorial, and a root menu with choices of four options: a picture naming task, a dialect quiz, a speech map in which free speech can be recorded, and an "about" section. In this paper, the data collected with the picture naming task and the dialect quiz are considered only. For an overview of all the functionalities of the application, the researchers' ethical considerations, and the collection of other speech and language attitude data, see Hilton (2021).

3.1   The Stimmen picture naming task

In the Stimmen application, a picture naming task can be found with 87 hand-drawn images of everyday objects in the Netherlands. The scientific purpose of this module is to collect data that can be used to document phonological and phonetic patterns for any language recorded, but the nameability and selection with regards to avoiding heteronyms was done on the basis of the varieties spoken in Fryslân, only. The 87 constructs represent all the phonemes and their allophonic realisations in the most spoken varieties found in the region were the activities surrounding Stimmen were primarily conducted: in the north of the Netherlands. The current list of languages in which one can record includes 38 different languages, but more can be added by contacting the Stimmen team, or the author of this paper.

When opening the picture task from the root menu a prompt informs the participants that the aim of the task is to name as many pictures as possible, and in as many languages as they can. Next, participants are asked to fill in a short questionnaire about where they are from (to indicate this on a map), their gender, their age bracket, which languages they are most fluent in, which languages they actively use in their life, and to answer the open question whether there is anything they would like to share with the researchers about their own variety. Upon considerations with the ethics committee and legal experts associated

with the project it was decided that further biographical data could not be collected due to privacy concerns, as the speech recordings are publicly available alongside the social information collected.
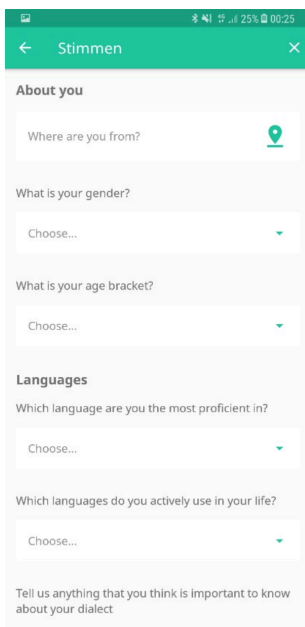

Figure 1: The meta-data questionnaire used for all tasks in 'Stimmen'.

Participants are then asked which language they would like to record in. After choosing a language, a randomised selection of a picture is made and shown to the user. They can then press the screen to record, and thereafter send the recording to the researchers. A prompt asks whether the user is sure they want to share the recording publicly. After naming 10 pictures the user can choose to go on or go to the gallery of named pictures to check their progress.

3.1.1 The recordings made in the picture naming task

Some 2,000 distinct individuals have used the picture naming task up until 2020, creating 41,553 individual recordings of words. 24,214 of these were created by female speakers, 17,028 by male speakers and 311 by speakers identifying as 'other'. The age distribution of the recordings is found in Figure 5. Almost half of the recordings are made by respondents younger than 30 years, which means that the recordings in Stimmen represent younger language. For the sake of investigating language change in progress this may not be a disadvantage, but for any
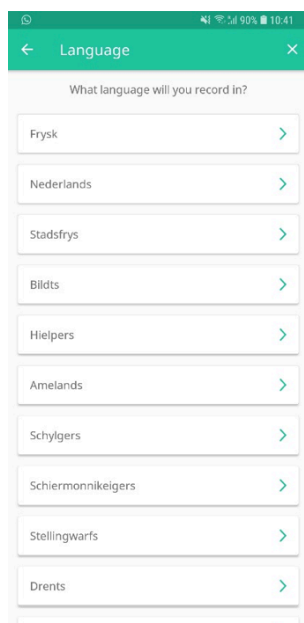
Figure 2: Some of the possible recording languages in 'Stimmen'.
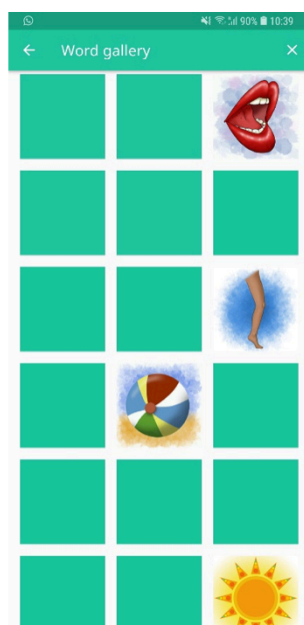

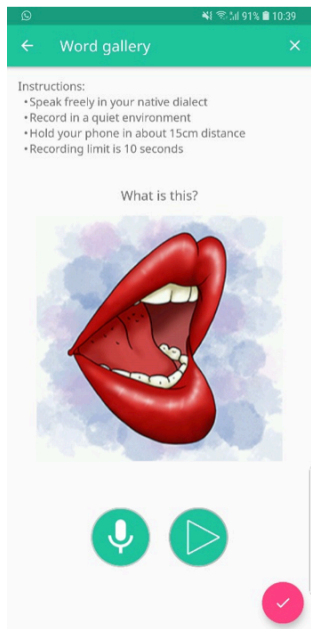Figure 3: Word gallery after naming four pictures in 'Stimmen'.

Figure 4: Screenshot of the picture task in 'Stimmen', showing the picture 'mouth'.

study reliant on balanced age data, one would need to take this imbalance into account in the analysis.
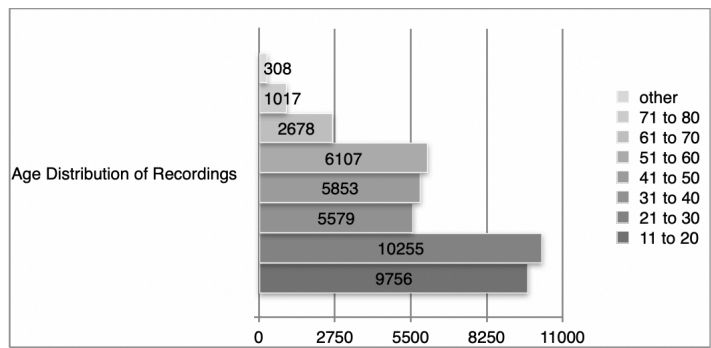

Figure 5: Distribution of recordings from the picture task in 'Stimmen' across age groups.

The top eight languages recorded in the picture naming task are shown in Table 1. Nearly three quarters of the recordings are made in Frisian, or varieties spoken within Fryslân.

Table 1: Number of Recordings Made by Languages in Stimmen Picture Task

Language	No. of recordings	Language	No. of recordings
Frisian	31,580	English	1,246
Dutch	5,858	German	207
Town Frisian	914	French	168
Bildts	540	Others	1,040

3.2 The Stimmen dialect quiz

A second component of the Stimmen application is a gamified task that guesses where people hail from within the Province of Fryslân. This task, "dialect quiz", was based on Leemann & Kolly (2013) and their subsequent applications. Data gathered with this module provides researchers with the public's knowledge of traditional linguistic variants, and regional and social (age, gender) variation in reported dialect use can be analyzed. A straight-forward way to develop a dialect quiz is to use an existing corpus of speech samples and create a set of questions that will result in a unique combination of answers for each of the locations represented in the corpus. Naturally, this is not easily done for minority languages. However, for the case of Frisian, mixed languages and Low Saxon as spoken in Fryslân previous dialectological work facilitated this. The prediction in Stimmen was created based on the GTRP database (2017) with data from 58 informants from the 1980s in Fryslân. The GTRP allowed for creation of a multilingual prediction for 58 different locations in Fryslân in which varieties of Frisian, Low Saxon and mixed languages are spoken.

While the picture naming task described in 3.1 was developed for use in *any* language, this part of the application was directed only at inhabitants of Fryslân. The quiz game asked users to provide their local variant for 19 Standard Dutch words (see Table 1), as all inhabitants of Fryslân are believed to be fluent in Dutch. 2-10 possible variants were available to the user to listen to and to choose between. After giving their personal variant to all 19 words the app makes three guesses of where a user could be from. The user can then indicate whether this prediction is correct or not and fill in the same meta-data questionnaire as used for the picture naming task (see Figure 1).

Table 2: Words in Stimmen's Dialect Quiz

Standard Dutch prompt	Translation	No. of variants
armen	'arms'	9
avond	'evening'	9
bij	'bee'	10
blad	'leaf'	3
borst	'breast'	8
dag	'day'	4
deurtje	'door' (+ diminutive)	12
geel	'yellow'	5
(ik ben) gegaan	'I have gone'	11
(ik heb) gezet	'I have placed'	6
heel	'whole'	5
kaas	'cheese'	7
koken	'cook' (infinitive)	11
oog	'eye'	7
(toe)sprak	'spoke'	8
tand	'tooth'	8
trein	'train'	5
vis	'fish'	2
zaterdag	'Saturday'	7

3.2.1 The data collected in the dialect quiz component

Of the 15,131 times the dialect quiz has been used, the survey has been filled in 3,340 times: 1,688 times by a female user; 1,633 by a male user and 19 times by a user identifying as 'other'. This rather low proportion (some 22%) of respondents filling in meta data could have to do with the fact that the survey was easily skipped in the popular web version of the Quiz (it was the very last component and the respondents had already received their guessed location). The age distribution of those using the Dialect Quiz and filling in the survey is given in Figure 6. Note that, as was the case in the picture task recordings, the youngest generations are over-represented in the collected data.

The project website houses interactive maps of the distributions of the answers for all variables and all variants, on http://stimmen.nl/uitspraakkaarten/.
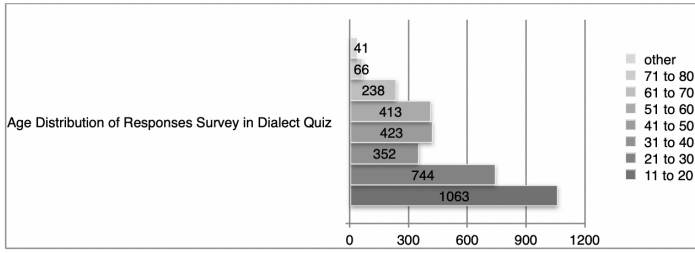
Figure 6: Distribution of survey responses in the Dialect Quiz in Stimmen across ages.

4 Analysis of variation in the Stimmen corpus

The manner chosen to validate the data collected with the picture naming task and the dialect quiz in the Stimmen project is to compare patterns of variation in the data with that found in studies collected with different methodologies, in approximately the same time period, in the Province of Fryslân. There are three comparisons that can be made with data in the Stimmen corpus: i) a comparison between the variation patterns in production of the coda cluster (sk) in the picture naming task, alongside reported usage of the variants from the dialect quiz data, with usage patterns found by Hilton & Weening (2014), ii) a comparison between the reported usage of the pronunciation variants for the Frisian word for 'eye' *each* in Stimmen with reported usage of the variants in Stefan, Klinkenberg & Versloot (2014), and iii) a comparison between the variation patterns in (r) in the data from the picture naming task with that in Van Bezooijen (2009).

4.1 Variation in (sk)

The coda cluster (sk) in Frisian has two variants: [sk] is the canonical Frisian variant but variant [s] is also commonplace. [s] is the variant used in equivalent cognates in the majority language Dutch. The variable occurs in two morphological contexts: (1) as the coda to a number of nouns and verbs that had <sk> as (part of the) coda in Old Frisian and other older Germanic varieties (OED Online, 2015); i.e., *fisk* 'fish'; *wask* 'wash'; *bosk* 'forest' or, more frequently throughout the lexicon, as (2) the coda in the adjectival or adverbial suffix equivalent to English 'ic': a) *histoarysk* 'historic' *fantastysk* 'fantastic', in which it occurs in the loan-morpheme −ysk from High German and Dutch. The Dutch equivalent phoneme in cognate words in both (1) and (2) above is /s/ cf. Dutch *vis*; *was*; *bos; historisch*; *fantastisch.* Hilton & Weening (2014) find lexical constraints exist on

the variation with certain adjectives and nouns less likely pronounced with the full cluster, and some words, such as *gânsk* 'whole' never undergoing variation. Furthermore, a higher writing proficiency in the minority language correlates with a higher proportion of the minority variant [sk] in speech. Hilton & Weening (2014) further argue that the variant [sk] is associated with being authentically Frisian. The data from Hilton & Weening (2014) has been made available for comparison with the data collected in the Stimmen app. That data was collected using a translation task and a 'map task' (where two participants must lead each other through incongruent maps) with 31 participants aged 15-62 (M=39.9, SD=14.2); where 15 participants were male, 16 female.

In the Stimmen application, the lexical item 'fisk' exists both in the dialect quiz, where 3144 of the informants who filled in questionnaires reported their variant usage, as well as in the picture naming task, where 419 recordings were made. 14 recordings were discarded from the corpus, due to the recording being inaudible, or the naming of the picture as a particular breed of fish (most often *bears* 'perch'). In the data from Hilton & Weening (2014), 76 tokens of 'fish' exist. The distribution of the pronunciation variants of 'fish' in the three different data sets are rendered in Table 3–5. The Dialect Quiz and Picture Task data render the same distributions as the data collected by Hilton & Weening (2014). It is also clear that female speakers use a higher proportion of [s] than males in all three data sets.

Table 3: The distribution of variants [s] and [sk] in the three data sets compared

	Stimmen dialect quiz data	
	female	male
[sk] (%)	1118 (70.7%)	1123 (72.8%)
[s] (%)	464 (29.3%)	420 (27.2%)
Total	1582 (100%)	1543 (100%)

4.2 Variation in (each) - 'eye'

A second linguistic variable in the crowd sourced Stimmen data that can be compared to other research is the reported pronunciation of 'eye' *each*. This linguistic variable has a standardised variant in written Frisian, which is representative

Table 4: The distribution of variants [s] and [sk] in the three data sets compared

| | Stimmen picture task data | |
	female	male
[sk] (%)	184 (76.3%)	135 (82.3%)
[s] (%)	57 (23.7%)	29 (17.7%)
Total	241 (100%)	164 (100%)

Table 5: The distribution of variants [s] and [sk] in the three data sets compared

| | Hilton & Weening (2014) | |
	female	male
[sk] (%)	28 (71.8%)	29 (78.4%)
[s] (%)	11 (28.2%)	8 (21.6%)
Total	39 (100%)	37 (100%)

of one of three main variants that exist in spoken language [ɪ.əx], while variants [e:x] and [ɛ:x] are associated with particular regional varieties of Frisian (cf. Breuker 1992: 19), making the variable particularly suitable to consider when assessing whether regional dialect levelling, or convergence, is ongoing in the Frisian speech community.

No published studies of this variable exist, but the authors of an on-going linguistic survey project (Stefan, Klinkenberg & Versloot p.c.) in Fryslân have kindly provided a map with distributions of the three Frisian variants for *eye* in 250 survey responses. The study uses a traditional dialectological methodology with a survey including direct questions about language use and attitudes (see Stefan et al., 2014 for more details of the methodology). The resulting dialectological map, shown in Figure 7, can be compared to the dialect maps made with the answers from the dialect quiz in the Stimmen project.

In figure 7, the regional distributions of the variants for eye in the comparison dataset is shown. The variant is indicated by the colour of the municipality as well as in the pie charts. The western and southern municipalities as shown in Figure 8 have a majority of *eech* [e:x], while the central north and north-west has
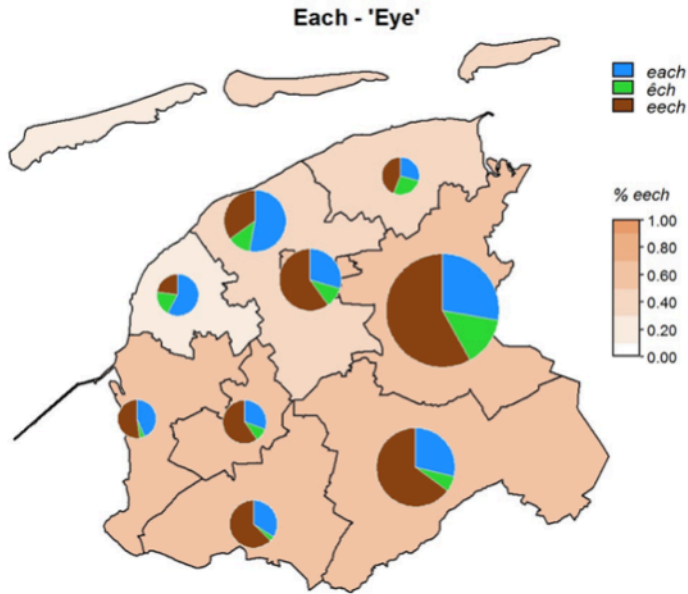
Figure 7: Distribution of the variants for 'eye' in an ongoing study by Stefan, Klinkenberg & Versloot (p.c.).

a majority of standard variant *each* [ɪ.əx]. The final variant êch [ɛ:x], indicated with green colour, is particularly frequent in the north-eastern municipality. If looking at distribution maps[1] for variant usage in the Stimmen corpus in Figures 8–10, we see a very comparable picture in the distribution of the variants (despite using lower level neighbourhood borders, as opposed to municipal borders). Each map contains the number of realisations of the different variants in the top right corner. Each realisation is a unique user. The brown variant in Figure 7, is *eech* [e:x] in Figure 8 and shows dark red colours throughout the southern areas and up alongside the western border of Fryslân. Variant *each* [ɪ.əx] (blue in Figure 7) variant has the highest proportion in the central and western part of the north of Fryslân also in Figure 9. Finally, the êch [ɛ:x] variant (green in Figure 7) shows the highest proportion in the north-eastern area in Figure 10. The dialect quiz data, then, shows a regional distribution of variants that follow the same pattern as that in findings of investigation using more traditional dialectological methodologies.

---

[1]I would like to extend a heartfelt thanks to Herbert Kruitbosch for creating the interactive maps of the responses in the dialect quiz. They can be found on stimmen.nl/uitspraakkaarten
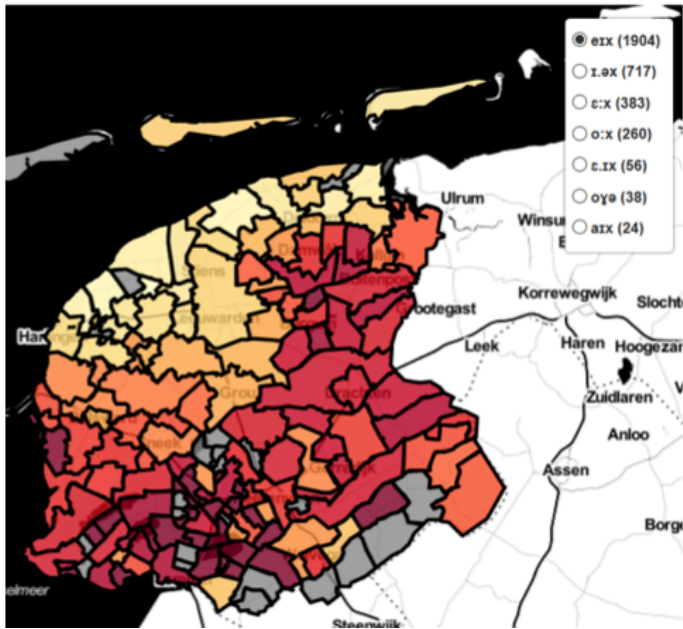
Figure 8: The regional distribution of the eech [e:x] response in the Stimmen Dialect Quiz, corresponding to the brown variant in Figure 7. Darker colours indicate higher reported usage frequency, grey areas indicate no reported usage.

4.3 Variation in (r)

The last sociolinguistic variable for which we can compare previous data with that from the Stimmen project is (r). A highly noticeable sound change that has taken place in Dutch in recent decades has been the innovation and spread of variant [ɹ], the approximant bunched realisation of post-vocalic coda /r/. This is a feature popularly referred to as a 'Gooise r'. In a large-scale study Sebregts (2015) concludes that age, region as well as gender predict the use of the approximant variant in Dutch, with female speakers using the highest proportion of the [ɹ], a finding supported by evidence in Van Bezooijen (2005). One interesting research question is whether, in highly bilingual populations, sociolinguistic variants are taken on also in a second, closely related language. This is what Van Bezooijen (2009) considers, studying pronunciation in a picture naming task eliciting coda /r/. In the data collected from 26 Frisian-speaking and 30 Town-Frisian subjects, there is no sign of convergence towards variation patterns found in Dutch. No approximant codas are found at all in her analysis of the realisations of 'boer' *farmer* and 'gieter' *water can*. The main variant of /r/ used in Fryslân is alveolar, with uvular variants present in respondents who speak Town Frisian.
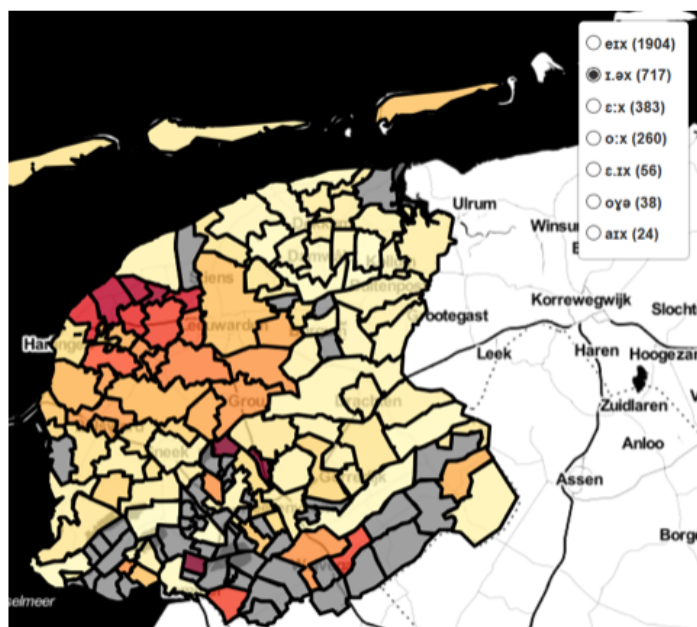
Figure 9: The regional distribution of the each [ɪ.əx] response in the Stimmen Dialect Quiz, corresponding to the blue variant in Figure 7. Darker colours indicate higher reported usage frequency, grey areas indicate no reported usage.

Bakker (2018) uses the spoken data collected in the Stimmen project labelled 'Frisian' to consider whether [ɹ] has gained ground in Fryslân a decade after Van Bezooijen (2009). Using auditory analysis and an additional coder he considers variation in the recordings of 'ear' *ear* from the Stimmen picture task and concludes that the usage of the approximant variant is very rare (1.6% of the cases), as can be seen in Table 3–3. He calls for more research, as he finds that the approximant variants are produced by young female speakers and discusses whether the Stimmen data could show the first stage of introduction of the variant to Frisian. For the purpose of the aim of the current paper, however, it is fair to conclude that a percentage of 1.6% is comparable to the finding that Van Bezooijen made, of negligible usage of the approximant variant in coda position in Frisian speech.

Note that Bakker (2018) attests a larger proportion of uvular variants (7.3%) in the Frisian Stimmen data than that attested in Van Bezooijen (2009; 0%). The difference between the two findings might be explained by the fact that the Stimmen data includes recordings from a much larger regional area than that included by Van Bezooijen (2009). In the Stimmen data, respondents living in areas in which
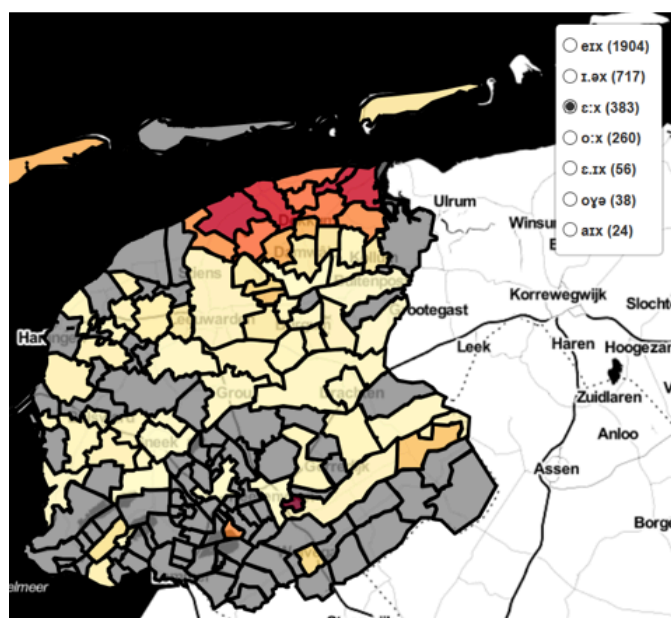
Figure 10: The regional distribution of the êch [ɛ:x] response in the Stimmen Dialect Quiz, corresponding to the green variant in Figure 7. Darker colours indicate higher reported usage frequency, grey areas indicate no reported usage.

Table 6: Variants of /r/ produced (as coded by Bakker 2018) in the Stimmen picture naming task ('ear'), and in the data reported by Van Bezooijen (2009).

	Bakker (2018) 'ear' in Stimmen data	Van Bezooijen (2009) 'boer' and 'gieter' in Frisian	Van Bezooijen (2009) 'boer' and 'gieter' in Town Frisian
alveolar	323 (87.5%)	52 (100%)	48 (80%)
uvular	27 (7.3%)	0 (0%)	12 (20%)
approximant	6 (1.6%)	0 (0%)	0 (0%)
'other'	13 (3.6%)	0 (0%)	0 (0%)
Total	369 (100%)	52 (100%)	60 (100%)

Town Frisian and Bildts are spoken (and where the uvular variant is widespread) feature in the data set. This could possibly have resulted in a higher proportion of uvular /r/ recordings, not only for the data labelled as "Town Frisian", but also for the data labelled as "Frisian". Bakker (2018) concludes that the most uvular variants in the Stimmen data are produced by subjects who state they come from Leeuwarden, the Sneek, or Franeker areas, which are all areas in which Town Frisian is spoken alongside Frisian and Dutch. The development of Town Frisian and its relationship to Frisian on a phonological level is a topic worthy of further research, on the basis of the findings in Bakker (2018).

5 Discussion and conclusion

The aim of this paper has been to validate the use of crowd sourced language data from minority languages for linguistic research. Crowd sourcing is a possible means to increase available research data from such languages. Currently the amount of available data for research from such languages is minimal, something that was also true for the minoritised varieties in the north of the Netherlands before the initiation of the Stimmen project. Overall, Stimmen has been a successful attempt in gathering large amounts of speech data from minority language users, but primarily those in the Netherlands. The data collected in the project is predominantly from younger users but is well-balanced in terms of gender representation.

Previous considerations of validity of crowd sourced data indicates that quality can be comparable to that collected with traditional methods, as long as the task for the user is not too complex. In the Stimmen application, users were asked to record their own words for pictures in the Picture Naming Task, or to indicate which variant they use in a Dialect Quiz, made especially for inhabitants of Fryslân. When considering validity of a method one should ideally test the outcomes from the methodology against a larger test battery. However, the amount of research data available from minority communities, including the community in question in this paper, Frisian, gives a rather small basis of comparison. Yet, the relatively recent sociolinguistic work that is available for comparison, Hilton & Weening (2014), Stefan, Klinkenberg, and Versloot (fc.) and Van Bezooijen (2009) all display very similar results to those found in analyzes of the data from Stimmen. This indicates that the data collected through the Stimmen application is valuable resource for investigating linguistic variation.

The finding that crowd sourced data from minority languages can be used for research of variation in speech is welcome. In a world in which travel, and

contact, restrictions have become more common-place the opportunity to collect data remotely is a game changer. Mobile phones and other portable technology that allows contact with a speech community within a simple click should be taken into use as much as possible. The inclusion of data collected using smartphones and virtual games can lead to much wider representation, not only of different linguistic varieties, but also of more participants, in research. Future work in this field should focus on whether remote efforts also increases the participation of contested language users, such as new speakers, and how one can ensure that a representative sample of all age groups is included in remote studies.

References

Bakker, Wiljo. 2018. *The way you /r/.* Groningen: University of Groningen. ().

Bonney, Rick, Tina B. Phillips, Heidi L. Ballard & Jody W. Enck. 2016. Can citizen science enhance public understanding of science? *Public Understanding of Science* 25(1). 2–16.

Breuker, Pieter. 1992. Taalkundige skaaimerken fan normearre frysk. *Us Wurk* 41(1-2). 1–58.

Brunelle, Marc. 2009. Diglossia and monosyllabization in eastern cham: A sociolinguistic study. *Variation in indigenous minority languages.* 47–75.

Centraal Bureau voor de Statistiek. 2018. *CBS StatLine - regionale kerncijfers Nederland.* https : / / statline . cbs . nl / Statweb / publication / ?DM = SLNL&PA = 70072ned&D1 = 0 - 88 , 292 - 293&D2 = 6 , 20 - 22&D3 = 21 - 23&VW = T (1 January, 2019).

De Decker, Paul & Jennifer Nycz. 2011. For the record: Which digital media can be used for sociophonetic analysis? *University of Pennsylvania Working Papers in Linguistics* 17(2). 51–59.

De Haan, Germen J. 1997. Contact-induced changes in modern west frisian. *Us Wurk* 46(1-4). 61–89.

De Vries, Nic J., Marelie H. Davel, Jaco Badenhorst, Willem D. Basson, Febe De Wet, Etienne Barnard & Alta De Waal. 2014. A smartphone-based asr data collection tool for under-resourced languages. *Speech communication* 56. 119–131.

Dorian, Nancy. 1978. The fate of morphological complexity in language death: Evidence from East Gaelic. *Language* 54(3). 590–609.

Feitsma, Tony, Els van der Geest, Frits van der Kuip & Irénke Meekma. 1987. Variations and development in frisian sandhi phenomena.

Goeman, Ton, Johan Taeldeman & Piet van Reenen. 2017. *Mand/fand/gtrp-database, dialecttranscripties 1980–1995*. http://www.meertens.knaw.nl/mand/database/.

Hilton, Nanna Haug. 2021. Stimmen: A citizen science approach to minority language sociolinguistics. *Linguistics Vanguard* 7(s1). 1–15.

Hilton, Nanna Haug & Joke Weening. 2014. Poster presented at sociolinguistics circle – university of groningen. In *Is this standardisation? Variation in (sk) in Frisian.*

Hoekstra, Erik & Marjo van Koppen. 2000. Het bildts als resultaat van fries-hollands taalcontact. *Amsterdamer Beiträge zur Älteren Germanistik* 54. 89.

Horn, Alexander. 2019. Can the online crowd match real expert judgments? how task complexity and coder location affect the validity of crowd-coded data. *European Journal of Political Research* 58(1). 236–247.

Kolly, Marie-José & Adrian Leemann. 2015. Dialäkt äpp: Communicating dialectology to the public–crowdsourcing dialects from the public. *Trends in phonetics and phonology. Studies from German-speaking Europe.* 271–285.

Leemann, Adrian & Marie-José Kolly. 2013. *Dialäkt äpp*. https://itunes.%20apple.%20com/ch/app/dialakt-app/id606559705 (16 January, 2014).

Leemann, Adrian, Marie-José Kolly, Ross Purves, David Britain & Elvira Glaser. 2016. Crowdsourcing language change with smartphone applications. *PloS one* 11(1). 1–25.

Lind, Fabienne, Maria Gruber & Hajo G. Boomgaarden. 2017. Content analysis by the crowd: Assessing the usability of crowdsourcing for coding latent constructs. *Communication methods and measures* 11(3). 191–209.

Möller, Robert & Stephan Elspaß. 2015. *21. atlas zur deutschen alltagssprache (ada).*

Nagy, Naomi. 2009. The challenges of less commonly studied languages. *Variation in indigenous minority languages, Impact: Studies in language and society* 25. 397–417.

Nagy, Naomi & Miriam Meyerhoff. 2008. Poster presented at NWAV2008. In *The love that dare not speak its name: The fascination with monolingual speech.*

Nicholas, Joe. 1988. British language diversity surveys (1977–87): A critical examination. *Language and Education* 2(1). 15–33.

Noglo, Kossi. 2009. Sociophonetic variation in urban Ewe. *Variation in indigenous minority languages.* 229–244.

Nota, Amber, Nanna Haug Hilton & Matt Coler. 2016. Word and phrasal stress disentangled: Pitch peak alignment in Frisian and Dutch declarative structures. *Speech Prosody 2016.* 464–468.

O'Rourke, Bernadette & Fernando Ramallo. 2013. Competing ideologies of linguistic authority amongst new speakers in contemporary galicia. *Language in Society* 42(3). 287–305.

O'Shannessy, Carmel. 2009. Language variation and change in a north australian indigenous community. *Variationist approaches to indigenous minority languages.* 419–439.

Sebregts, Koen. 2015. *The sociophonetics and phonology of Dutch r.* Leiden: LOT.

Shameem, Nikhat. 1998. Validating self-reported language proficiency by testing performance in an immigrant fijians. *Language Testing* 15(1). 86–108.

Shing, Han-Chin, Suraj Nair, Ayah Zirikly, Meir Friedenberg, Hal Daumé III & Philip Resnik. 2018. Expert, crowdsourced, and machine assessment of suicide risk via online postings. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, 25–36.

Sjölin, Bo. 1976. *"min Frysk": Een onderzoek naar het ontstaan van transfer en" code-switching" in gesproken Fries.* Frysk Ynstitut oan de Ryksuniversiteit to Grins.

Stanford, James N. 2016. A call for more diverse sources of data: Variationist approaches in non- English contexts. *Journal of Sociolinguistics* 20. 525–541.

Stanford, James N. & Dennis Preston. 2009. *Variation in indigenous minority languages* (IMPACT: Studies in Language, Culture and Society 25). Amsterdam: John Benjamins Publishing.

Steemers, Patrick, Marc Beijn, Dirk Reuser, Remco Gaykema, Erik Bookholt, Joel van Veen & Mike Odenhoven. 2017. *Deloitte global mobile consumer survey 2017: The Dutch edition.* (2 July, 2018).

Stefan, Nika, Edwin Klinkenberg, Arjen Pieter Versloot, et al. 2014. Frisian sociological language survey goes linguistic: Introduction to a new research component. *Philologia Frisica Anno.* 240–257.

TaalAtlas. 2015. *De Fryske taalatlas 2015, Fryske taal yn byld.* Leeuwarden.

Van Bezooijen, Renee. 2005. Approximant/r/in Dutch: Routes and feelings. *Speech communication* 47(1-2). 15–31.

Van Bezooijen, Renee. 2009. The pronunciation of/r/in frisian. *Variation in Indigenous Minority Languages* 25. 299.

van Bree, Cor. 1994. Het probleem van het ontstaan van het 'stadsfries' in verband met nieuwe talen in contact-theorieën. In *Predota, stanislaw (red.). handelingen regionaal colloquium neerlandicum wroclaw*, vol. 1993, 43–66.

Vaux, Bert. 2004. Let's go usa: 2004. In chap. American dialects. Let's Go Publications.

Weinreich, Uriel, William Labov & Marvin Herzog. 1968. *Empirical foundations for a theory of language change.* Vol. 58. University of Texas Press Austin.

Wenker, Georg. 1881. *Sprach-Atlas von Nord-und Mitteldeutschland: Text und Einleitung: Auf Grund von systematisch mit Hülfe der Volksschullehrer gesammeltem Material aus circa 30000 Orten.* Strassburg: Trübner.