Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
iNEST
unibz Freie Universität Bozen Libera Università di Bolzano Università Liedia de Bulsan
ALPILINK

Language Documentation via Crowdsourcing

Birgit Alber
in cooperation with Joachim Kokkelmans and Emily Siviero
birgit.alber@unibz.it
Free University of Bozen/Bolzano

Workshop on Language Diversity, Data Elicitation and Multilingualism
Leibniz-ZAS, Berlin, 26 May 2023

Overview

- New ways of studying linguistic diversity: crowdsourcing
- The project *VinKo-AlpiLink*
  - Features of VinKo-AlpiLink (lights and shadows)
  - Quality of the data
  - Quantity of the data
  - Citizen science projects (*Vinkiamo*)

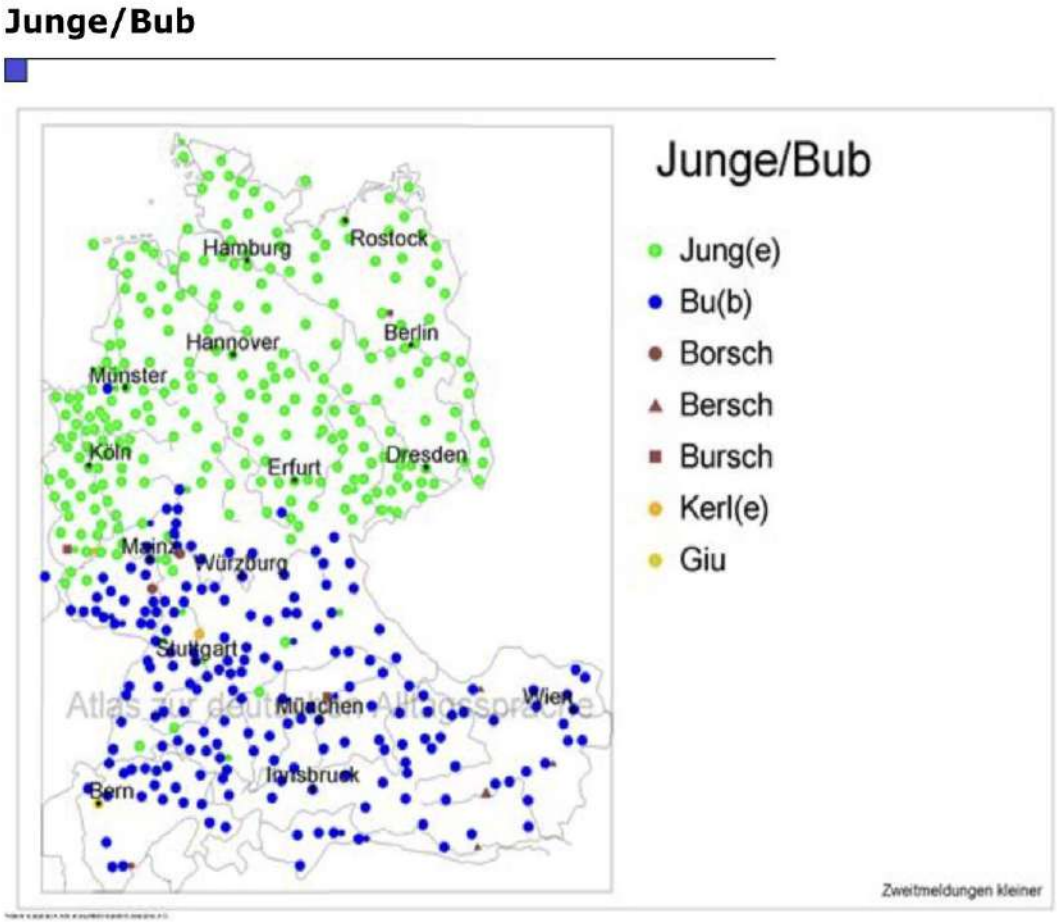
Crowdsourcing

A new way of studying language diversity: crowdsourcing

AdA: Atlas zur deutschen Alltagssprache

http://www.atlas-alltagssprache.de/

- *Fragerunden* (question rounds)
- Instructions: 'Bitte geben Sie bei den folgenden Fragen jeweils an, welchen Ausdruck man in Ihrer Stadt normalerweise hören würde – egal, ob es mehr Mundart oder Hochdeutsch ist. Antworten Sie bitte, ohne lange nachzudenken!'

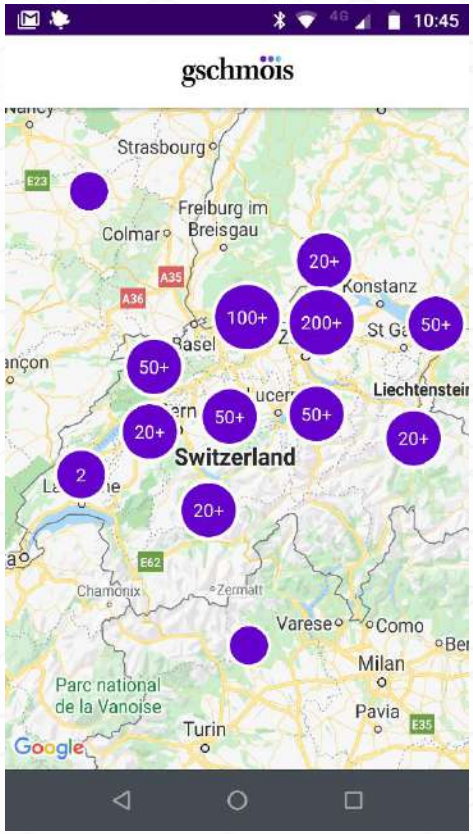


***Junge/Bub* (Frage 1)**

Crowdsourcing

A new way of studying language diversity: crowdsourcing

Gschmöis: Swiss Dialects https://www.gschmois.uzh.ch/de.html



Typically

- focus on the lexicon
- written responses

Crowdsourcing

Advantages

- Informants: many
- Data: a lot, geographically fine-grained
- Restitution: high potential of involvement of informants (*citizen science*); immediate restitution in form of maps or similar

But: Accompanied by a certain loss of control by the linguist/fieldworker →

Crowdsourcing

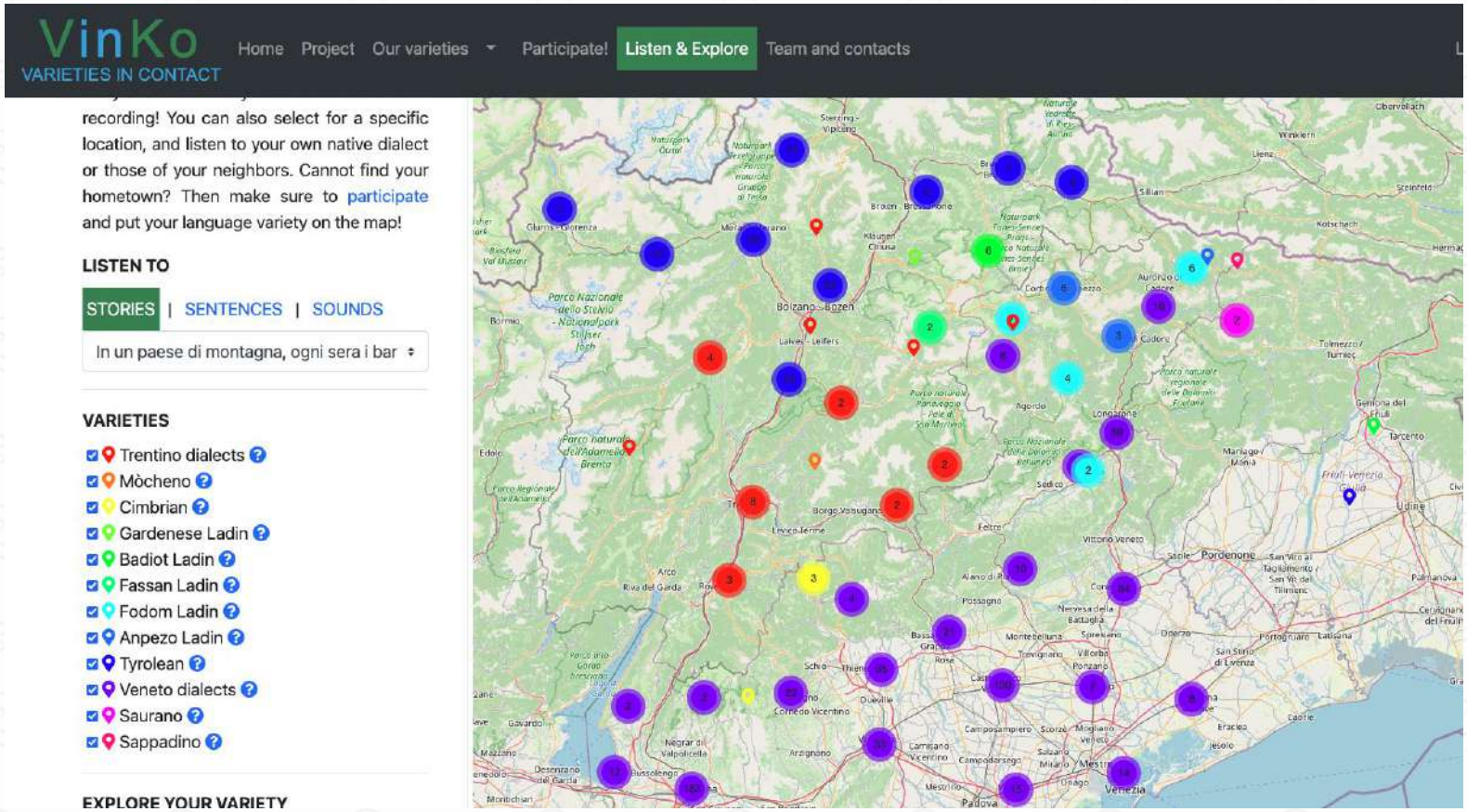
Critical points

- Which informants?
  - reliable speakers of the variety we want to elicit?
- Which data?
  - dialect, regional languages, regional standard?
  - form of the elicited data
    - written: can be used only for certain phenomena
    - audio data: technically challenging, may be noisy, has to be processed
    - impossibility of checking back with the informants
  - phenomena which can be studied: lexicon, phonology, morphology syntax?

VinKo - AlpiLink

Main Features (https://vinko.it/) (Cordin et al. 2018, Rabanus et al. 2022, Kruijt et al. b. in press)

- **Crosslinguistic data elicitation**: Germanic and Romance varieties of Northern Italy



VinKo - AlpiLink

- **Crosslinguistic data elicitation**: Germanic and Romance varieties of Northern Italy

Germanic	Romance
Tyrolean dialects	Trentino dialects
Mòcheno (lg. island)	Veneto dialects
Cimbrian (lg. island)	Ladin (Badia, Fassa, Fodom, Anpezo)
Saurano (lg. island)	*Piedmontese dialects*
Sappadino (lg. island)	*Franco-Provençal*
Walser (lg. island)	*Occitan*

Project partners:
Universities of

Verona: Stefan Rabanus (coordinator), Anne Kruijt
Bozen/Bolzano: Birgit Alber, Joachim Kokkelmans
Trento: Ermengildo Bidese
Torino: Livio Gaeta
Valle d'Aosta: Gianmario Raimondi

VinKo - AlpiLink

- **Crosslinguistic data elicitation**: Germanic and Romance varieties of Northern Italy

Same phenomena in different varieties:

- detection of contact-induced change
- possibility of typological comparisons

Example: Laryngeal features of stops

/p/ vs. /b/

Tyrolean:	*Paarl*	*Baam*	'type of bread, tree'
Veneto:	*pan*	*borsa*	'bread, bag'

VinKo - AlpiLink

Example: Laryngeal features of stops

- Vietti, Alber & Vogt (2018): acoustic analysis, Tyrolean dialect of Meran
    - 10 informants from Meran
    - laboratory setting (soundproof booth)
    - no aspiration in /p, t, k/
      [differently from more Northern German dialects]
    - categorical neutralization of labials /p ~ b/, word-initially
    - intraspeaker variability in neutralizing /k ~ g/ and /t ~ d/, word-initially

(1) a. [p]aarl 'Paarlbrot, type of bread"

b. [p]aam 'Baum, tree'

VinKo - AlpiLink

Example: Laryngeal features of stops

Qs

- can these results be replicated with a wider range of informants?
- is the contrast in terms of VOICE (not spread glottis) similar to that of the neighboring Romance languages?

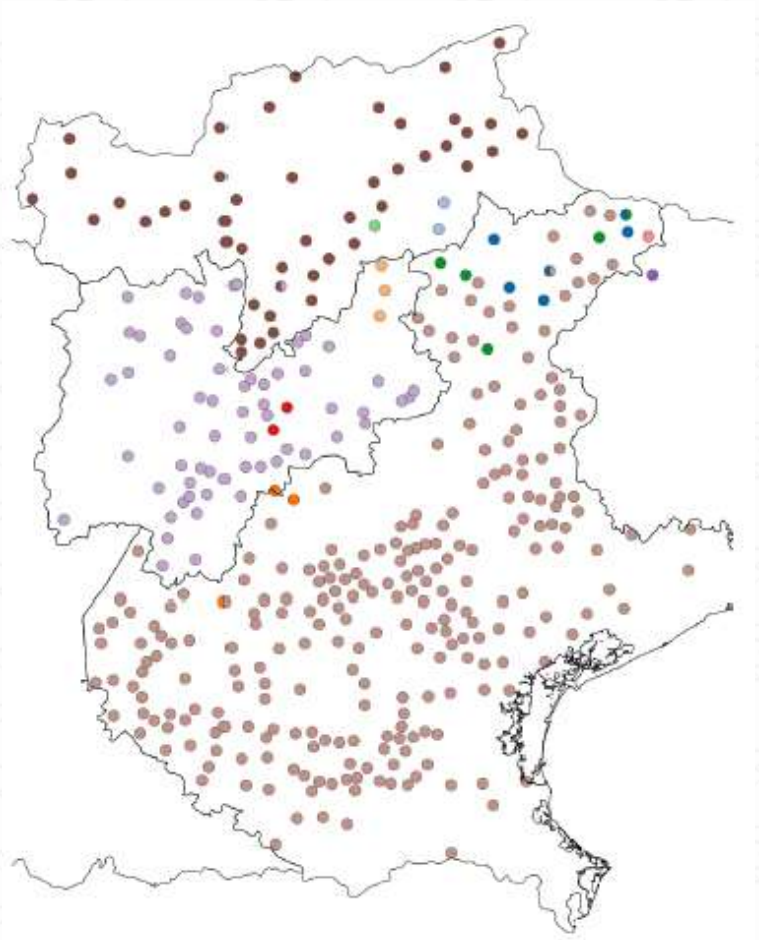
These are questions we hope to answer with the help of crowdsourced data of the VinKo-Alplilink project

VinKo - AlpiLink

- **Data over five-year period (2018-2022) (Kruijt 2023, Kruijt et al., in press b)**

- 1392 questionnaires
- each questionnaire: 152 to 181 audio recordings.
- more than 125.000 audio files
- 377 locations
- corpus accessible at Clarin repository, EURAC

http://hdl.handle.net/20.500.12124/46



VinKo - AlpiLink

- **Data over five-year period (2018-2022) (Kruijt 2023)**

data not necessarily well balanced: Tyrolean

	Average age	Gender M/F		Proficient	Frequent use	Use in family	Use with friends
Anpezan	47,5	42%	58%	92%	58%	67%	73%
Badiot	34,7	14%	86%	100%	86%	100%	100%
Cimbrian	56,3	71%	29%	86%	57%	43%	83%
Fassan	23,3	25%	75%	100%	100%	100%	100%
Fodom	64,3	43%	57%	100%	90%	95%	90%
Gardenese	39,5	0%	100%	100%	100%	100%	100%
Mòcheno	39,1	43%	57%	100%	100%	100%	100%
Sappadino	50,0	0%	100%	100%	100%	100%	0%
Saurano	67,5	100%	0%	50%	50%	50%	50%
Trentino	39,3	56%	44%	89%	84%	86%	64%
Tyrolean	26,0	10%	90%	99%	95%	97%	99%
Venetan	48,2	45%	55%	82%	72%	79%	76%
Overall	45,3	43%	57%	85%	76%	82%	78%

Table 5: Speaker information per language variety (17 January 2023)

VinKo - AlpiLink

- **Phenomena investigated: (synchronic) structure**

Examples:

Phonology	obstruent system
	processes involving sibilants
	/r/
Morphology	pronoun and determiner paradigms
	truncation patterns
Syntax	pro-drop
	complementizers/subordination
	particle verbs

VinKo - AlpiLink

- **Form of elicitation**

Phonology: word-list reading; separate questionnaires for each variety (with glosses)

Advantages: low impact of standard varieties

Critical points: ad hoc orthographies; sometimes items are not recognized

B71 | Wies

	A	B	C	D	E	F
33	32	Paarl	Paarlbrot	obstr	wi	/p/
34	33	Baam	Baum	obstr	wi	<b>
35	34	Toog	Tag	obstr	wi	/t/
36	35	Dåch	Dach	obstr	wi	/d/
37	36	Kårtn	Karten	obstr	wi	/kx/
38	37	Ggåggele	Ei	obstr	wi	/k/
39	38	Ggugger	Fernglas	obstr	wi	/k/
40	39	Gårtn	Garten	obstr	wi	/g/
41	40	oper	schneefrei	obstr	wm	/p/
42	41	låppet	dumm	obstr	wm	/p/

Ampezzano | Badiot | Cimbrian (Lusern) | Fassan | Fodom | Gardenese | Mòcheno | Trentino | Tyrolean | Veneto

VinKo - AlpiLink

- **Form of elicitation**

Morphology: personal pronoun paradigms and articles (Kruijt 2022)

- translation task aided by pictures embedded in a story
- guided free production
- text in standard language; prompting through audio examples in a Germanic/Romance dialect

Die Hexe will die gefangenen Kinder kochen

La strega vuole cucinare i bambini presi

VinKo - AlpiLink

- **Form of elicitation**

Syntax: translation task from standard language (German or Italian)

Subordination:

a. Ich weiß nicht, welcher Bus zuerst abfährt

b. Non so quale autobus parta per primo

'I don't know which bus leaves first'

Critical points

- influence from the standard languages?
- which standard language to choose as a trigger for varieties such as Mòcheno or the Ladin varieties?
- perceived as the most complex and boring task by informants (Kruijt 2023)

VinKo - AlpiLink

General critical points (Kruijt 2023)

- length of the questionnaire (70% of feedback to VinKo experience by informants)
- complexity of the task (reading a word list is easier than translating)
- phonological questionnaire: proposed words don't match words in the local dialect
- tasks too repetitive: stories might be interesting than the traditional tasks

VinKo – Quality of the Data

Is crowdsourced data reliable?

- Kruijt (2022): Subject clitics in Veneto dialects described in the literature are present in VinKo (but not in the standard trigger sentences)

Input:

Tornano di corsa in paese e chiedono aiuto al cacciatore

'[They] return immediately to the village and [they] ask the hunter for help'

Veneto dialect from Bonavigo (Kruijt 2022: 48):

I=	*torna*	*de corsa*	*in paese*	*e*	***i=***	*dimanda aiuto*	*a=l*
3PL.M=	return	in run	in village	and	3PL.M=	ask help	to-DEF.M.SG
cacciatore.							
hunter(M)							

More examples (morphosyntax) for the reliability of crowdsourced data: Rabanus (in press); Kruijt et al. (in press a.)

VinKo – Quality of the Data

Is crowdsourced data reliable?

- Alber & Kokkelmans (2022): **rhotics /r/** in Tyrolean dialects
- Comparison of three sources
  - Tirolischer Sprachatlas: 1965-1970
  - Scheutz (2016): data elicited in 2010
  - VinKo: data elicited from 2017-22
- Results
  - similar structures are found
  - where structures diverge, this can be explained by the diachronic change that rhotics are undergoing

VinKo – Quality of the Data

Is crowdsourced data reliable? (Alber & Kokkelmans 2022)

- Comparability of data: Scheutz (2016) – VinKo (2017-22)
  - age ('young speakers' in Scheutz – VinKo mean of 24 years)
  - social background (non-agriculture-related job – university students)
  - location (points within reasonable travel time = 50 mins by bike)
  - sex (female: 63% in Scheutz – 91% in VinKo)

VinKo – Quality of the Data

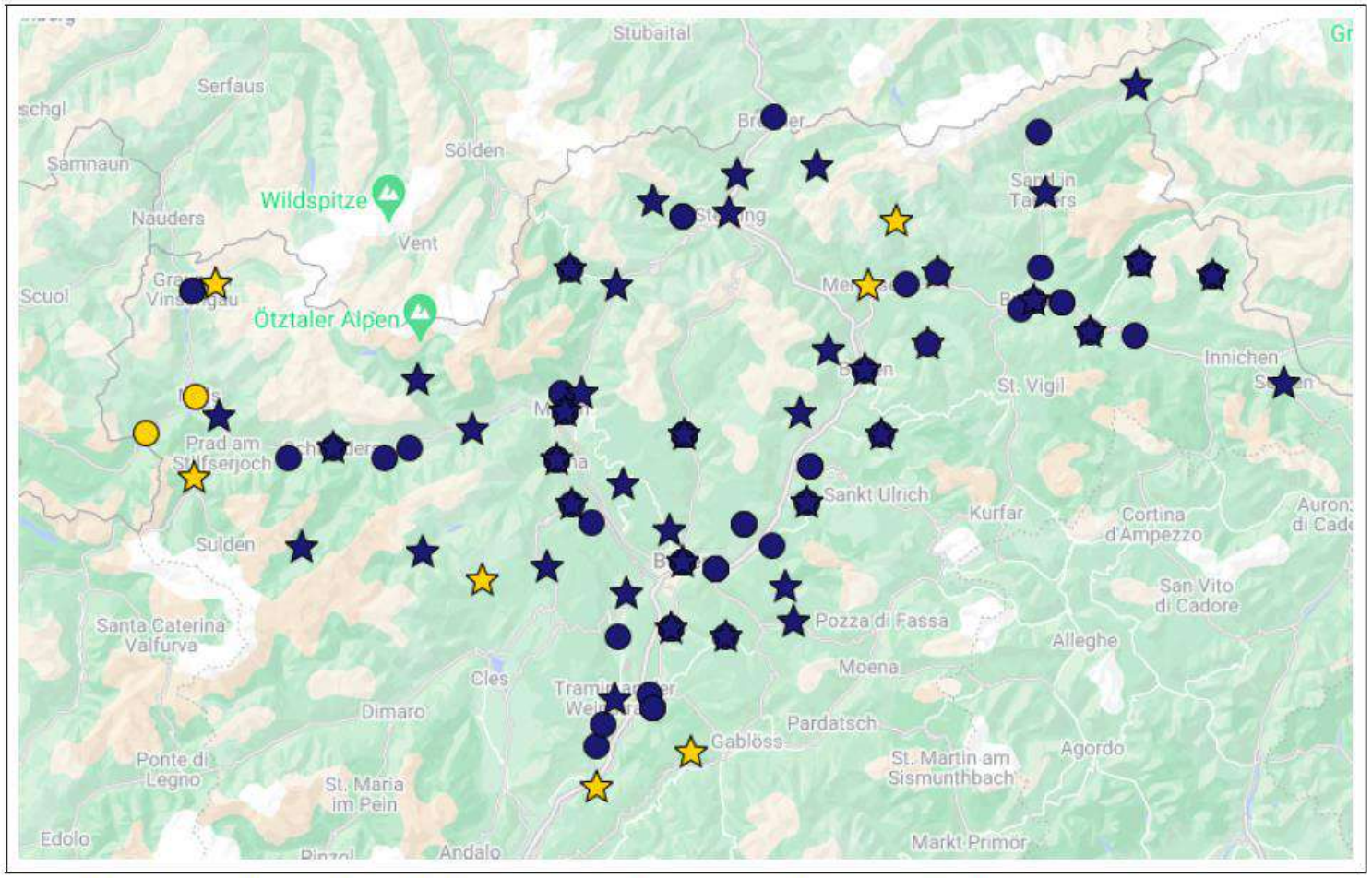
Is crowdsourced data reliable? (Alber & Kokkelmans 2022)

Rhotics in **Onset Position**: alveolar or uvular?

(2) lea.[r]e.[r]in vs. *lea.[ʀ]e.[ʀ]in* 'Lehrerin, teacher'

=> 84% match between Scheutz (2016) and VinKo (17-22)

VinKo – Quality of the Data



yellow = alveolar; blue = uvular; circles = *VinKo*; stars = *Insre Sproch*

VinKo – Quality of the Data

Is crowdsourced data reliable? (Alber & Kokkelmans 2022)

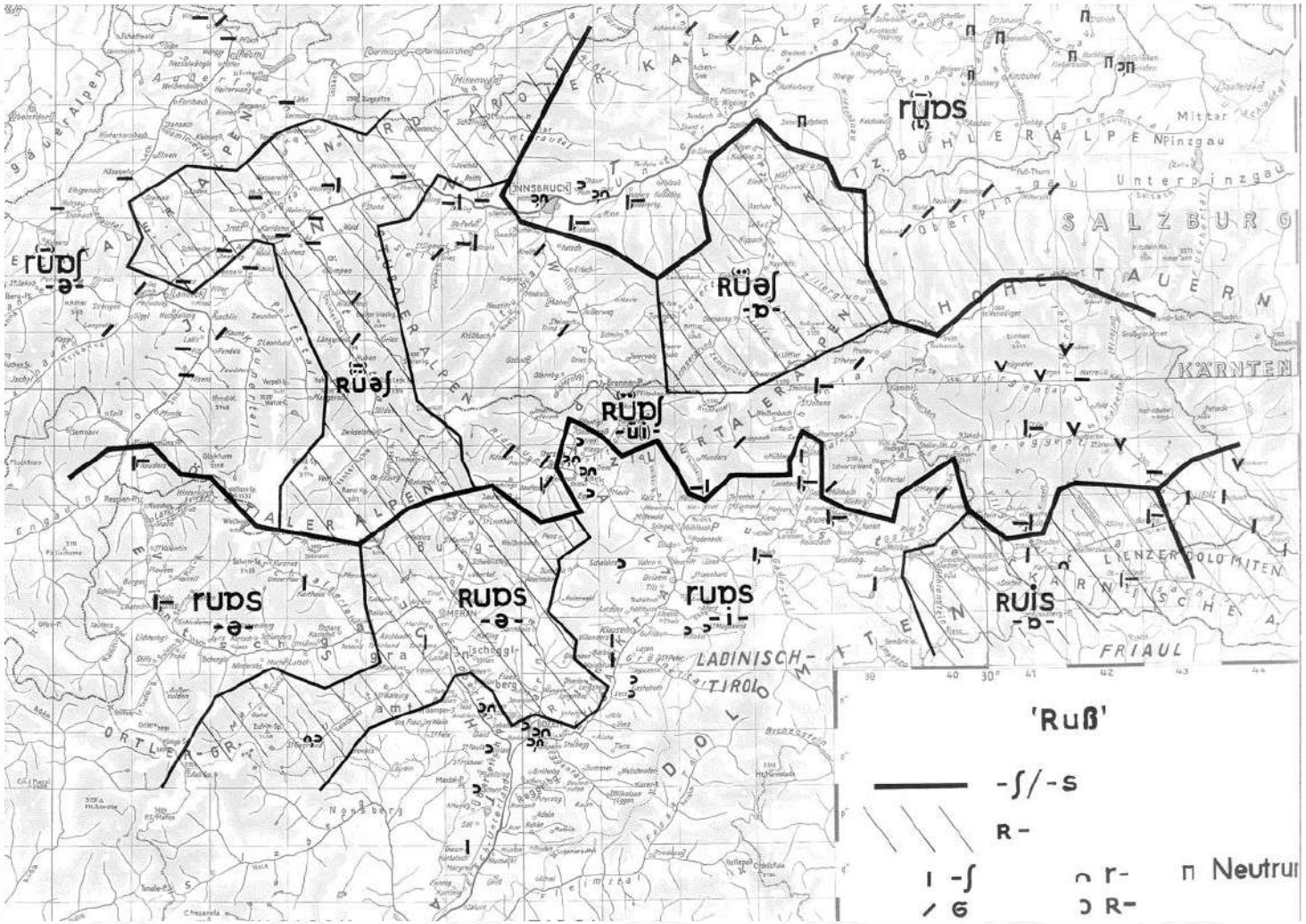
Interpretation

- 84% match for rhotics in **Onset position:** the diachronic shift from alveolar [r] to uvular [ʀ] is next to complete

Compare with older sources: Tirolischer Sprachatlas/Kranzmayer (1956)

- alveolar [r] most widespread variant
- uvular [ʀ] attested in urban contexts

VinKo – Quality of the Data



TSA

VinKo – Quality of the Data

Is crowdsourced data reliable? (Alber & Kokkelmans 2022)

Rhotics in **Coda Position**: consonantal or vocalized?

(3) Stressed syllables

a. after short vowel:	*du[r]/[ʀ]*	vs.	*du[ɐ̯]*	'dürr, meagre'
b. after long vowels:	*joo[r]/[ʀ]*	*vs.*	*joo[ɐ̯]*	'Jahr, year'

=> 68% and 66% match between Scheutz (2016) and VinKo (17-22)

VinKo – Quality of the Data

Is crowdsourced data reliable? (Alber & Kokkelmans 2022)

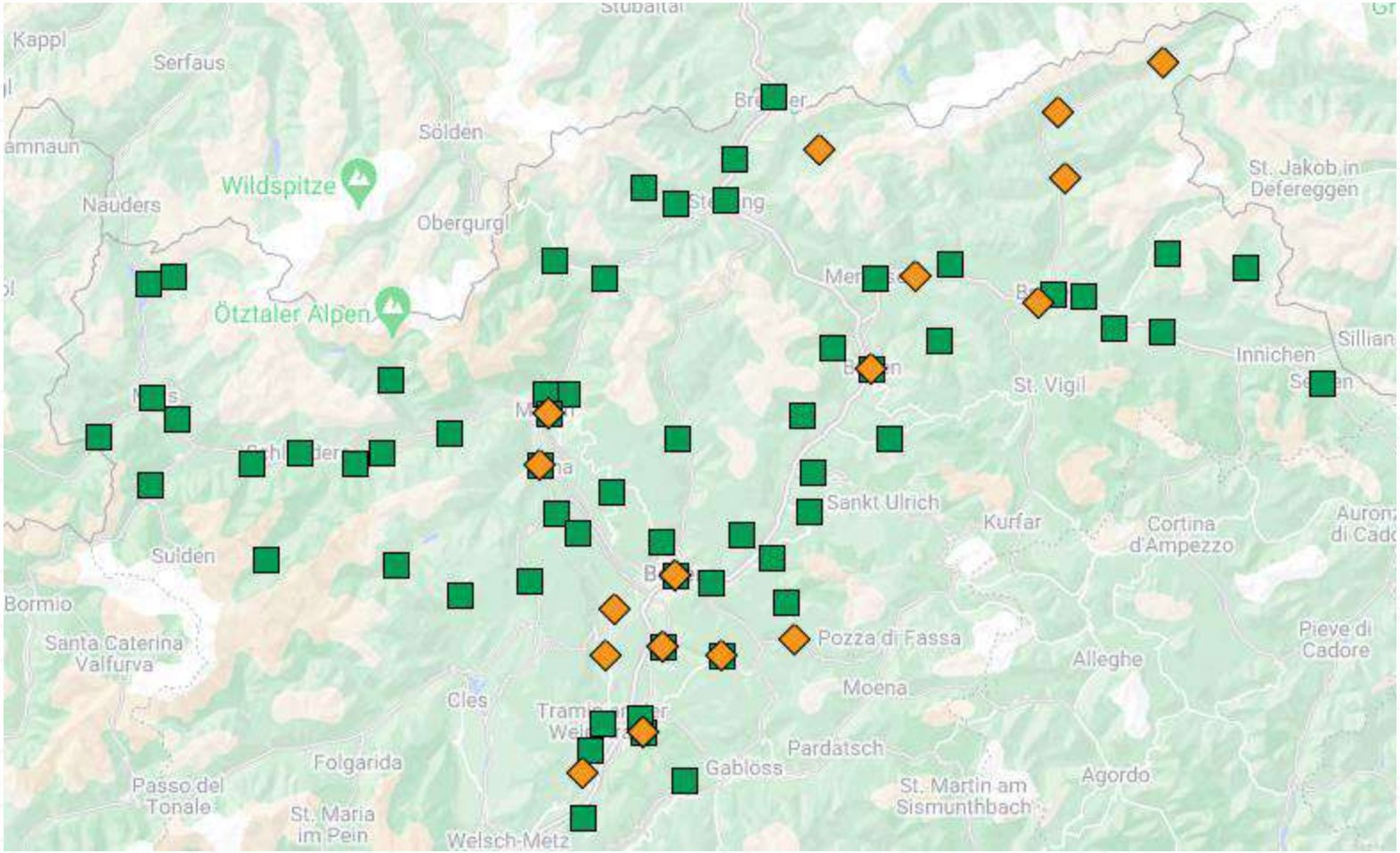
Rhotics in **Coda Position**: consonantal or vocalized?

(3) Unstressed syllables

wosse[r]/[ʀ] vs. *woss[ɐ]* 'Wasser, water'

=> 75% match between Scheutz (2016) and VinKo (17-22)

VinKo – Quality of the Data



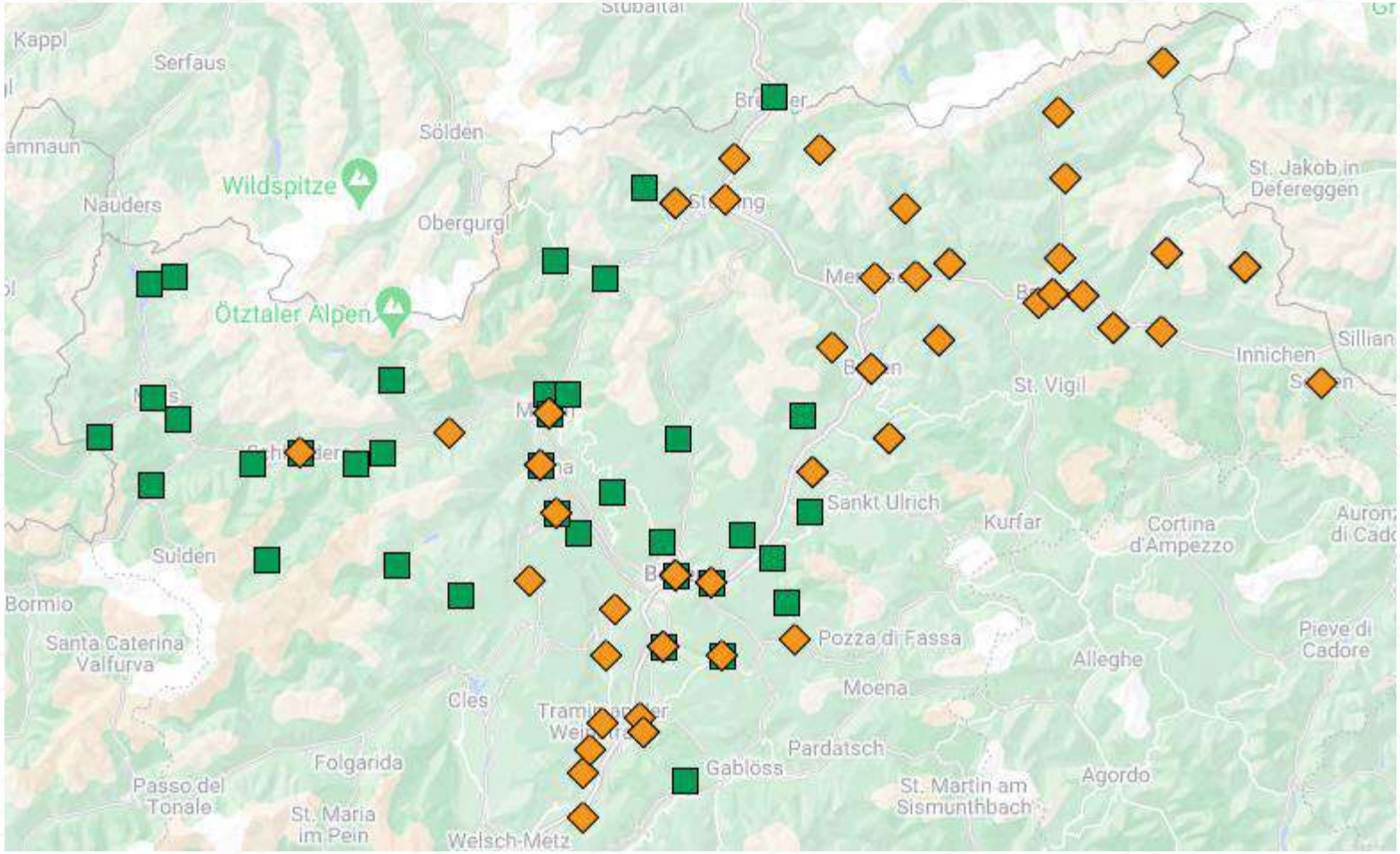
joor

green: consonantal

orange: vocalized

Combined data Scheutz (2016) + VinKo

VinKo – Quality of the Data



wosser

green: consonantal

orange: vocalized

combined data Scheutz (2016) + VinKo

VinKo – Quality of the Data

Is crowdsourced data reliable? (Alber & Kokkelmans 2022)

Interpretation

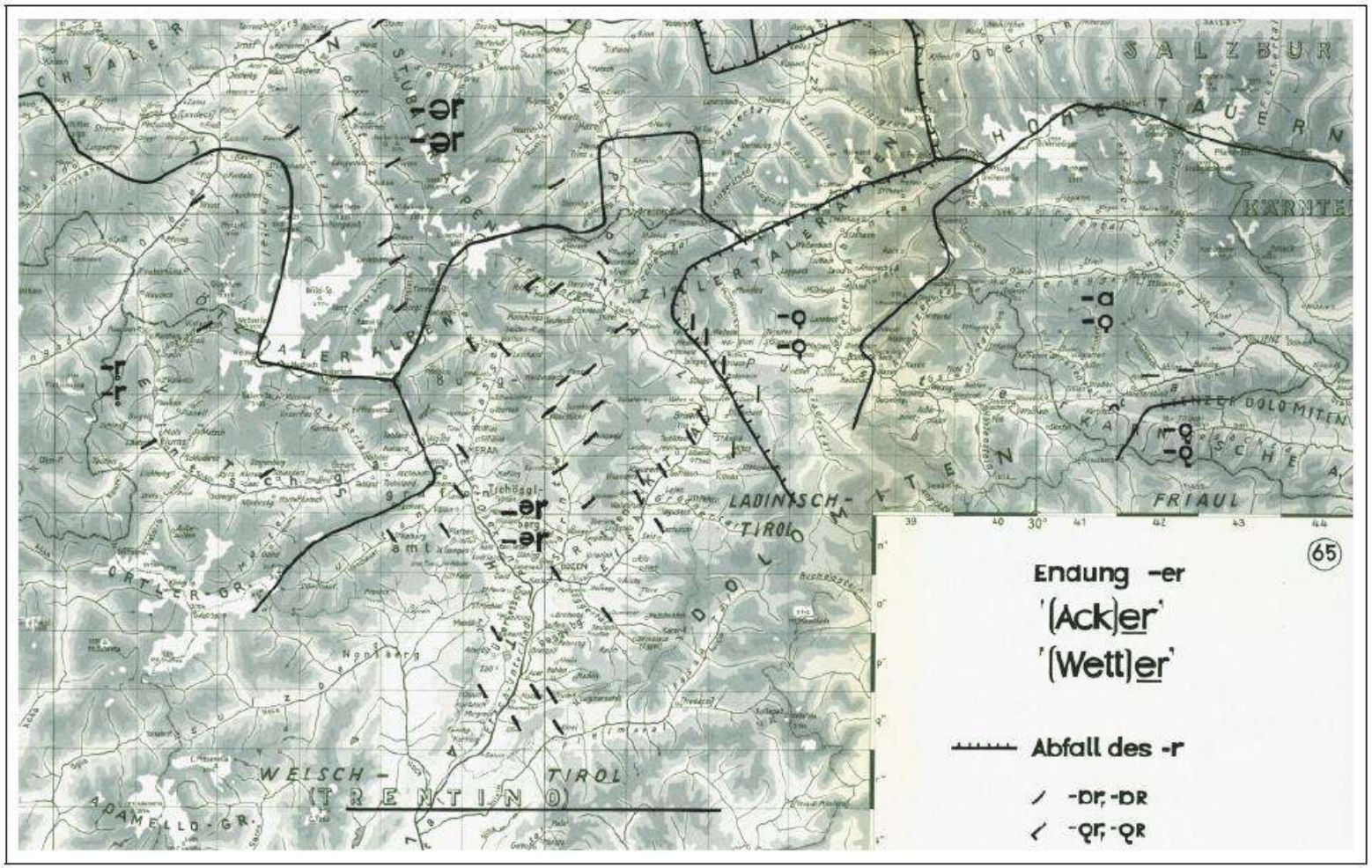
- 66-68-75% match for rhotics in **Coda position:** the vocalisation process is still ongoing, varying according to contexts

durr < *joor* < *wosser*

Compare with older sources: Tirolischer Sprachatlas

- vocalisation only attested for the *wosser*-Kontext, and only in the East (Pustertal)

VinKo – Quality of the Data



TSA

VinKo – Quality of the Data

Is crowdsourced data reliable?

YES!

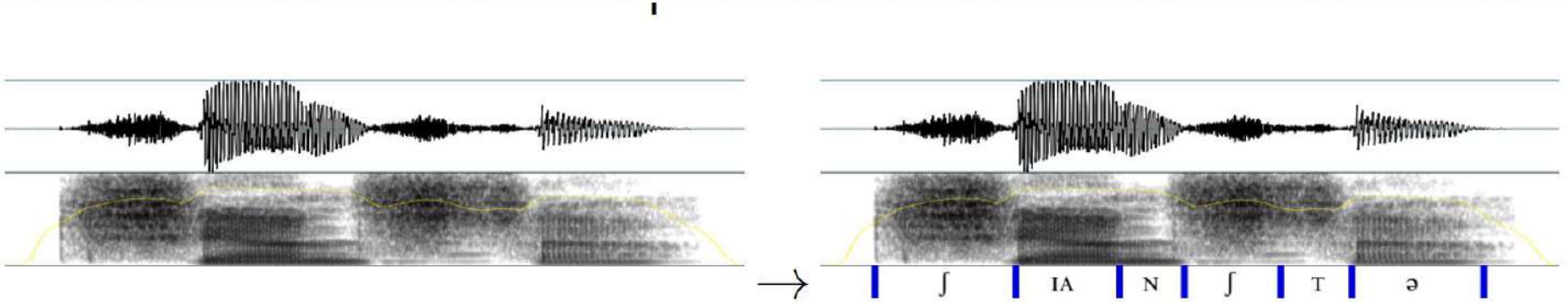
Comparison with 'traditional' sources shows that

- similar structures can be found in the data
- divergences are in line with observed paths of language change

VinKo – Quantity of the Data

A lot of data – how to deal with it?

From audio files to (possibly automatically) annotated corpus (Kokkelmans 2023)



Phonology questionnaire

- designed to elicit certain words → we know what we can expect (more or less)
- goal: isolate single segments to study their properties (e.g. laryngeal properties of stops)
- prepares the ground for future work: machine learning, transcription of sentences

VinKo – Quantity of the Data

A lot of data – how to deal with it? (Kokkelmans 2023)

possibly in an automated way, employing the least manual work

1. Data preprocessing
2. Forced alignment of segments with the help of WebMAUS (LMU Munich)
3. Manual correction of transcriptions
4. Development of interface for corpus search

VinKo – Quantity of the Data

A lot of data – how to deal with it? (Kokkelmans 2023)

1. Data preprocessing

Association of audio file of word-list item to three types of transcription

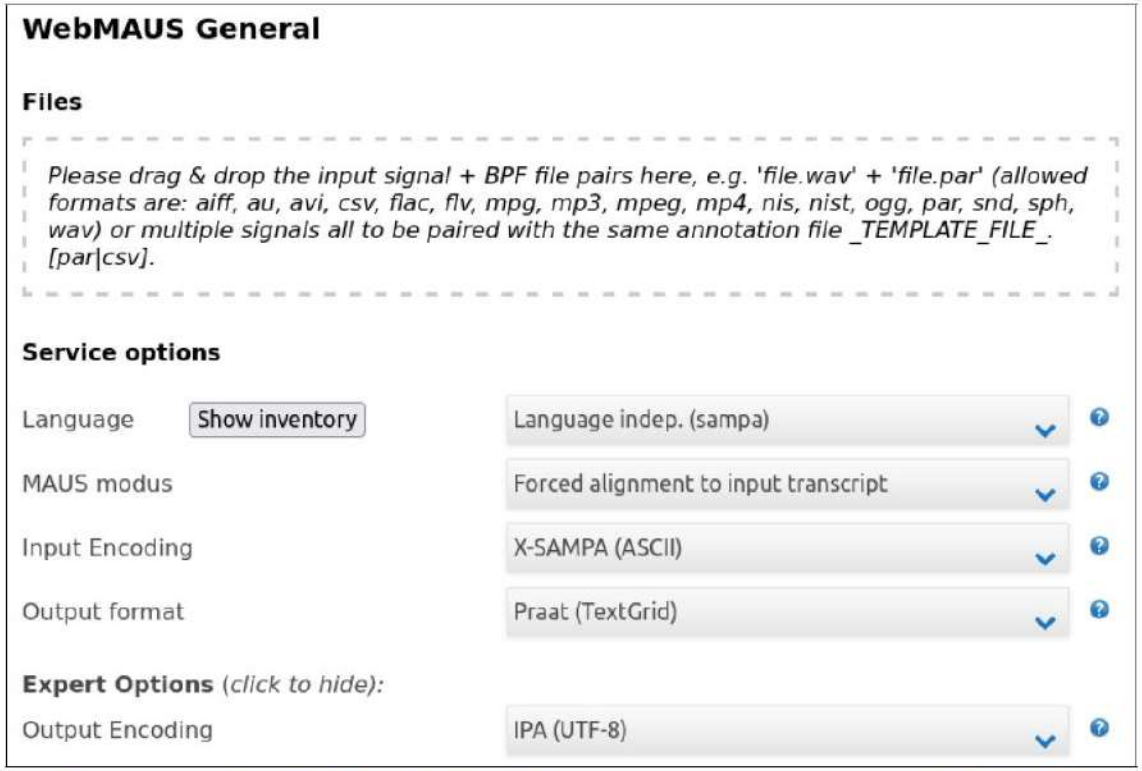
- VinKoGraphy: item as presented to informants
- HistPhonGraphy: transcription useful to follow variation in language change through dialects
- SAMPA: assumed phonetic realization (in majority of dialects)

	A	B	C	D	E
1	WordID	Lang	VinKoGraphy	HistPhonGraphy	SAMPA
2	W0001	tre	s e r v e l	k E r b E l	s E r v E l
3	W0002	tre	s a l t	g a l t	z a l t
4	W0003	tre	s a l	s` a l	s` a l

VinKo – Quantity of the Data

A lot of data – how to deal with it? (Kokkelmans 2023)

2. Forced alignment via WebMAUS

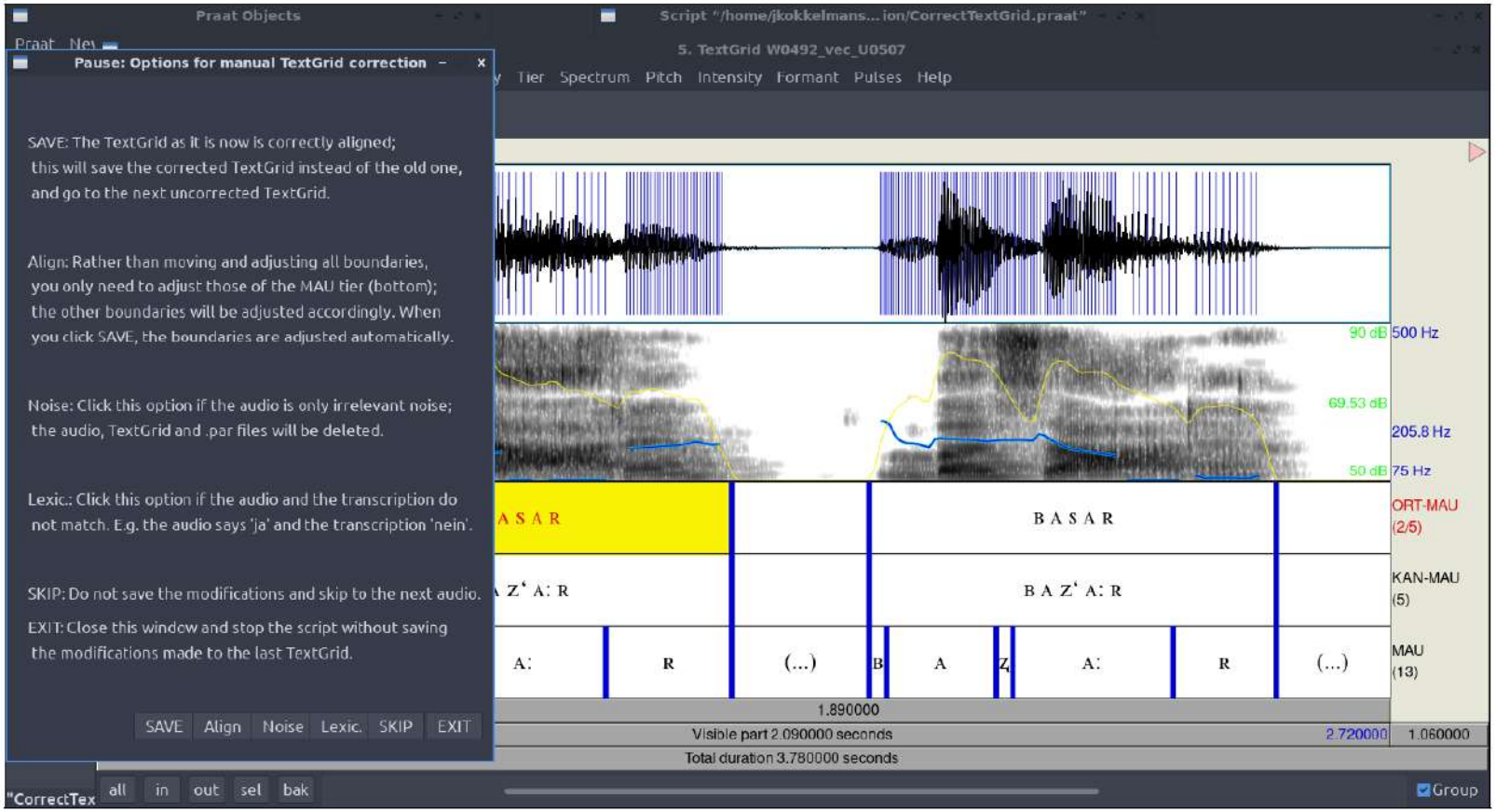


Schiel (1999, 2015); Kisler et al. (2017)

VinKo – Quantity of the Data

A lot of data – how to deal with it? (Kokkelmans 2023)

3. Manual correction



VinKo – Quantity of the Data

A lot of data – how to deal with it? (Kokkelmans 2023)

4. Interface for corpus search

here: <p> and <b> word-initially, in VinKoGraphy

Filter *VinKoQuestionnaires.csv* with regex

Regexes applied to filter the data:

Note: The option 'Match VinKoGraphy and SAMPA at the same place' is activated, meaning that the regexes for these two columns must refer to the same segment in the string.

WordID	Lang	VinKoGraphy	HistPhonGraphy	SAMPA	SAMPASAMPA
.*	.*	^(p|b|P|B)	.*	^(p|b|P|B)	.*

Generate list of files to analyse

Result of filtering:

Include?	WordID	Lang	VinKoGraphy	HistPhonGraphy	SAMPA	SAMPASAMPA	NumWords	Nb. of files
☑	W0005	tre	presa	prEs`a	prEs`a	prEs`a prEs`a	2	0
☑	W0011	tre	bus,busi	buk,buki	bus`,bu:z`i	bus`,bu:z`i bus`,bu:z`i	4	0
☑	W0012	tre	pesc,pesci	pisk,piski	pEs`,pEs`i	pEs`,pEs`i pEs`,pEs`i	4	0
☑	W0032	tre	bislonch	bis`long	biz`lONk	biz`lONk biz`lONk	2	0
☑	W0033	tre	pensar	pens`a:r	pEnz`a:r	pEnz`a:r pEnz`a:r	2	0
☑	W0042	mhn	bèsp	wEs`p	bEs`p	bEs`p bEs`p	2	3
☑	W0056	mhn	protestiarn	protEsti:r@n	prOtEs`ti:rn	prOtEs`ti:rn prOtEs`ti:rn	2	2
☑	W0065	mhn	birt	wirt	birt	birt birt	2	3
☑	W0088	cim	post	pOs`t	pOs`t	pOs`t pOs`t	2	4

VinKo – Quantity of the Data

A lot of data – how to deal with it? (Kokkelmans 2023)

Problems encountered on this path (corrected, in part, manually)

- Speakers don't follow instructions: 'read word twice'
- Speakers add comments: 'this word in my dialect really means y, not x ...'
- Speakers choose a lexical item different from the one proposed (*verruckt* instead of *norret* 'crazy')
- phonetic realization varies so much that WebMAUS has difficulties aligning:

Wirt → Wi[r]t, Wi[ʀ]t, Wi[ɐ]t, Wi[ʃ]t

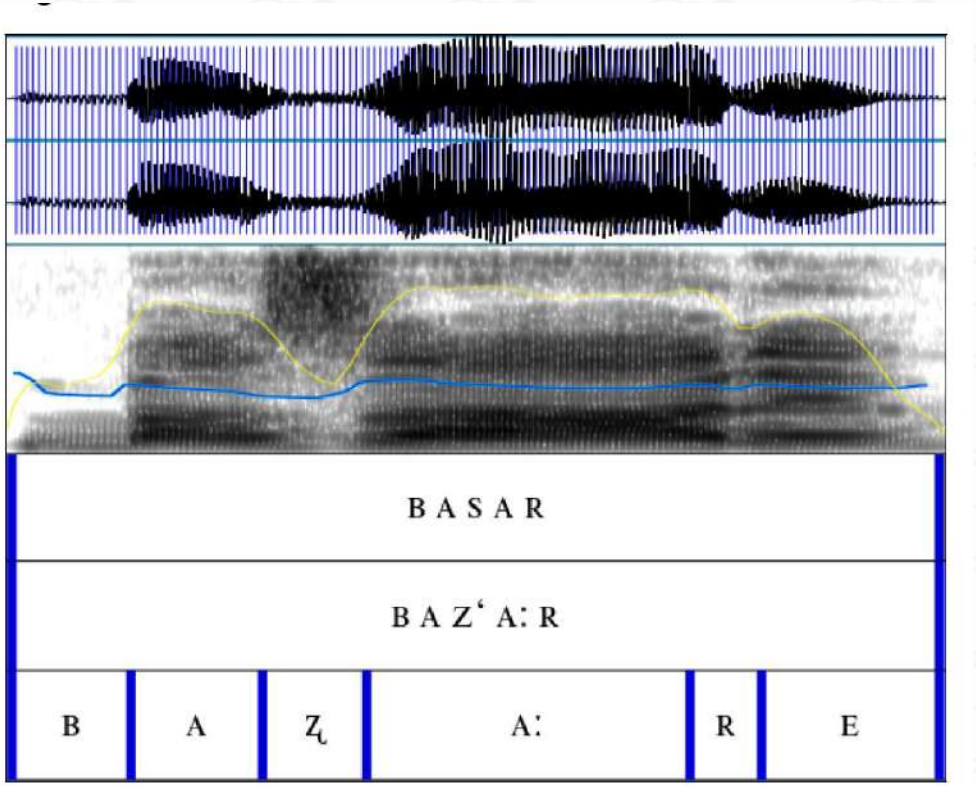
- some files consist only of noise or silence
- segments are added in some variety (differently from expected)

VinKo – Quantity of the Data

A lot of data – how to deal with it? (Kokkelmans 2023)

Problems encountered on this path (corrected, in part, manually)

- segments are added in some variety (differently from expected)



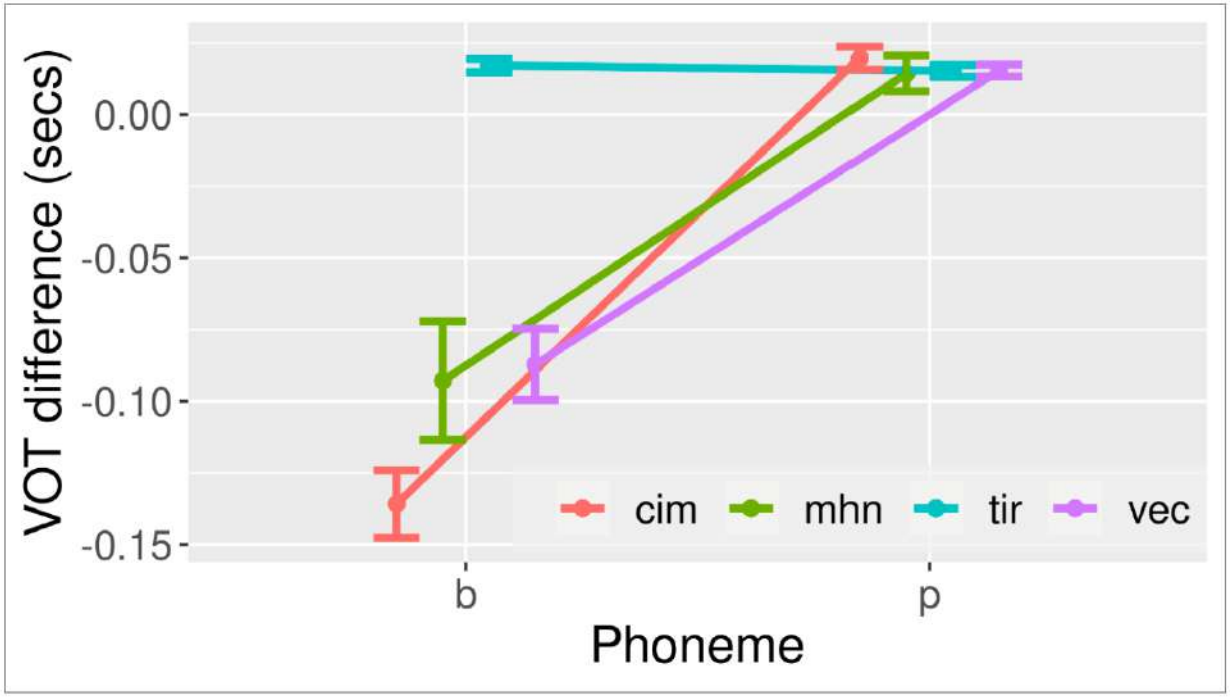
VinKo – Quantity of the Data

A lot of data – how to deal with it? (Kokkelmans 2023)

But eventually, analysis can start!

Here: historic /p/ vs. /b/ contrast, word-initially (118 items)

- Variable 1: *Voice Onset Timing*



First results

- neutralization of contrast in Tyrolean
- preservation of contrast in Cimbrian, Mocheno and the Veneto dialects

VinKiamo Südtirol

Crowdsourcing and Citizen Science

Crowdsourcing platforms have a high potential of involving the communities of speakers in activities related to language documentation

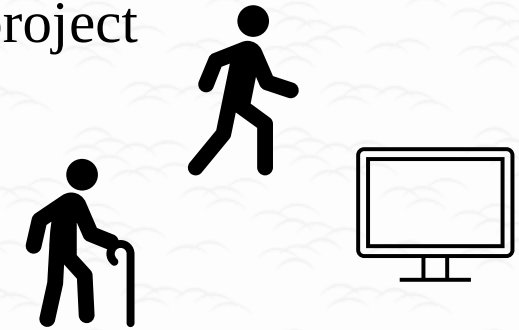
VinKiamo

- an offspring of VinKo-Alpilink, first implemented in Veneto in 2021 (University of Verona: https://sites.hss.univr.it/vinkiamo/)
- 2023 start of VinKiamo Südtirol (University of Bozen/Bolzano: Birgit Alber, Joachim Kokkelmans, Emily Siviero)

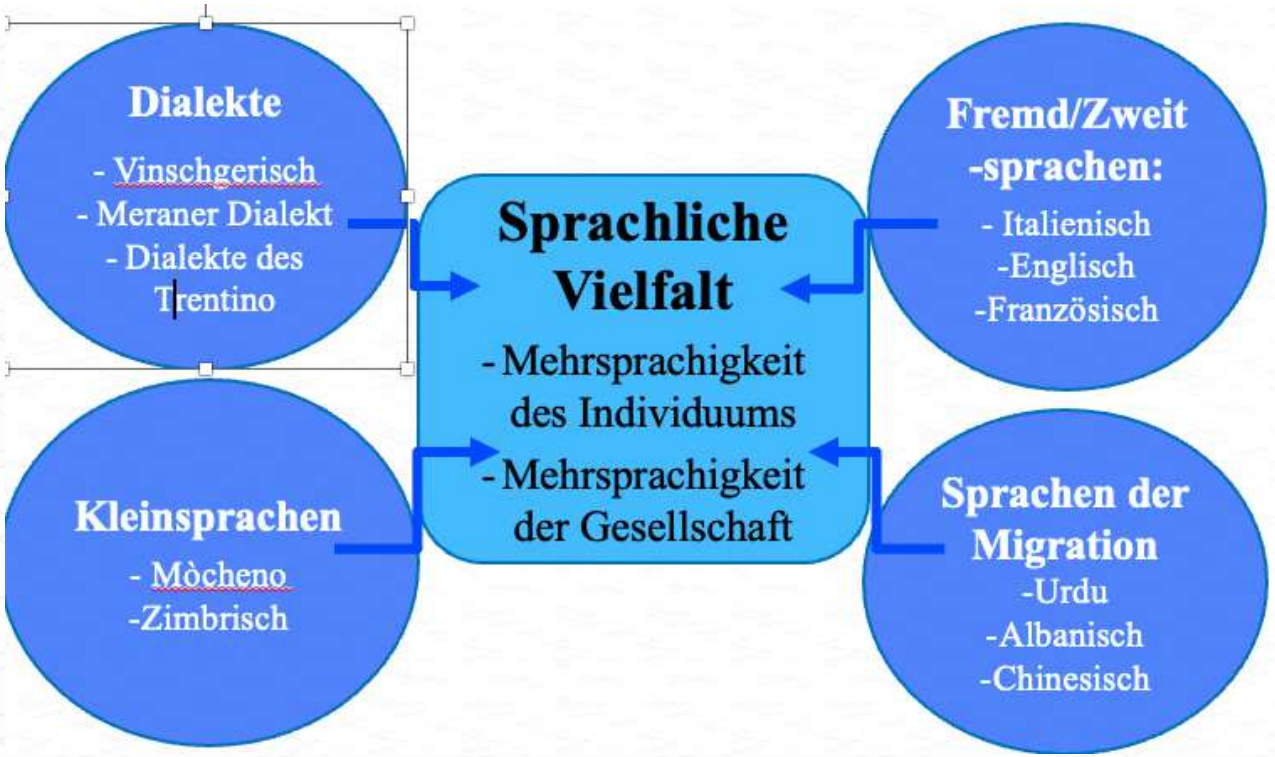
VinKiamo Südtirol

Structure

- target group: high-school students in South-Tyrol (about 17 years old)
- Presentation of basic concepts of linguistic diversity, multilingualism etc. [university staff]
- Introduction to the platform VinKo and to linguistic fieldwork [university staff]
- Fieldwork: students use VinKo to interview speakers of the older generation
- they write a report on their experience and leave a final overall feedback on the project



VinKiamo Südtirol



Was finden wir an Sprachenvielfalt schon alleine in diesem Klassenzimmer vor?
Welche Sprachen sprecht ihr?
Wir sammeln die Antworten an der Tafel ...

VinKiamo Südtirol

Goals

- Enhancing awareness of linguistic diversity
- Enhancing competences of students: organization of fieldwork, teamwork, data elicitation and analysis
- documenting the linguistic knowledge of the older generation
- bridging the digital gap between the younger and the older generation
- initiating an intergenerational discourse on linguistic diversity

Language Documentation via Crowdsourcing

Summary

- Language documentation via crowdsourcing: potentials and challenges
- The project VinKo-AlpiLink, with a focus on
  - elicitation of linguistic structure (phonology, morphology, syntax)
  - collecting data across the Germanic-Romance language family
  - collecting spoken language
- Potentials and challenges emerging in VinKo-AlpiLink
  - in data elicitation
  - concerning the quality of the data (no problem, so far)
  - concerning the quantity of the data
  - possibilities of citizen science projects

Acknowledgment of support by:

PRIN 2020 2020SYSYBS_002

ALPILINK

PNRR iNEST (Interconnected North-East Innovation Ecosystem)

I43C22000250006

ECS 00000043

Finanziato dall'Unione europea NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA

iNEST

unibz Freie Universität Bozen Libera Università di Bolzano Università Liedia de Bulsan

References

Alber, Birgit & Joachim Kokkelmans. 2022. South Bavarian rhotics in crowdsourced linguistic data from Northeastern Italy: a diachronic and qualitative comparison. Talk presented at the conference 'Beyond Borders: German-speaking Minorities in Italy and around the World', Trento, 6.10.2022.

Cordin, P., S. Rabanus, B. Alber, A. Mattei, J. Casalicchio, A. Tomaselli, E. Bidese & A. Padovan. 2018. VinKo, Versione 2 (20.12.2018, 09:20). In Lo spazio comunicativo dell'Italia e delle varietà italiane. Korpus im Text, 1–15. München: Ludwig-Maximilians-Universität

Kisler, T., U. Reichel & F. Schiel. 2017. Multilingul processing of speech via web services. *Computer Speech & Language* 45. 326-347.

Kokkelmans, Joachim. 2023. Halbautomatisierte Annotation von Dialektaufnahmen (VinKo & AlpiLink). Guest presentation at the Seminar 'Deutsche Dialektologie', MA in Applied Linguistics, Free University of Bozen/Bolzano.

Kranzmayr 1956 Kranzmayer, E. 1956. Historische Lautgeographie des gesamtbairischen Dialektraumes. Wien: Böhlau.

Kruijt, Anne. 2022. Crowdsourcing language contact: pronoun and article morphology in Trentino-South Tyrol and Veneto. PhD Dissertation: University of Verona.

Kruijt, Anne. 2023. VinKo. Final Report. Technical Report, University of Verona.

Kruijt, Anne, Patrizia Cordin, and Stefan Rabanus. (in press a.). 'On the validity of crowdsourced data'. In Elissa Pustka, Carmen Quijada Van den Berghe, and Verena Weiland (eds.): Corpus Dialectology: from methods to theory (French, Italian, Spanish). Amsterdam / Philadelphia: John Benjamins Publishing Company.

Kruijt, Anne, Stefan Rabanus, and Marta Tagliani (in press b.). 'The VinKo-Corpus: Oral data from Romance and Germanic local varieties of Northern Italy'. In Marc Kupietz and Thomas Schmidt (eds.): Neue Entwicklungen in der Korpuslandschaft der Germanistik: Beiträge zur IDS-Methodenmesse 2022. (= Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache (CLIP) 11). Tübingen: Narr.

Rabanus, Stefan (in press). Nome di battesimo e articolo espletivo – crowdsourcing e cartografica linguistica nello studio della variazione linguistica in Trentino-Alto Adige e Veneto. In: Schöntag, Rober & Laura Linzmeier (Hrsg.): Neue Ansätze und Perspektiven zur sprachlichen Raumkonzeption und Geolinguistik

Rabanus, Stefan; Kruijt, Anne; Tagliani, Marta; Tomaselli, Alessandra; Padovan, Andrea; Alber, Birgit; Cordin, Patrizia; Zamparelli, Roberto; Vogt, Barbara Maria. 2022, VinKo (Varieties in Contact) Corpus v1.1, Eurac Research CLARIN Centre, http://hdl.handle.net/20.500.12124/46.

Scheutz, H. 2016. Insre Sproch. Deutsche Dialekte in Südtirol. Mit dem ersten ”sprechenden” Dialektatlas auf CD-ROM. Bozen: Athesia.

Schiel, F. 1999. Automatic Phonetic Transcription of Non-Prompted Speech. In *Proceedings of the ICPhS*. San Francisco CA. 607-610.

Schiel, F. 2015. A Statistical Model for Predicting Pronunciation. In *Proceedings of the ICPhS*, Glasgow UK. Paper 195.

TSA = Klein, K. K., L. E. Schmitt & E. Kühebacher. 1965–1971. Tirolischer Sprachatlas (3 Bände) Deutscher Sprachatlas. Regionale Sprachatlanten Nr. 3: Tirolischer Sprachatlas. Marburg: N. G. Elwert Verlag.

Vietti, Alessandro, Birgit Alber & Barbara Vogt. 2018. Initial Laryngeal Neutralization in Tyrolean. Phonology. 35. 79-114