GeoLingIt at Evalita 2023

Overview of the Geolocation of Linguistic Variation in Italy Task

Alan Ramponi,[1] Camilla Casula[1,2]

[1]Fondazione Bruno Kessler [2]University of Trento

Evalita 2023 – *Sep 8th, 2023, Parma (IT)*

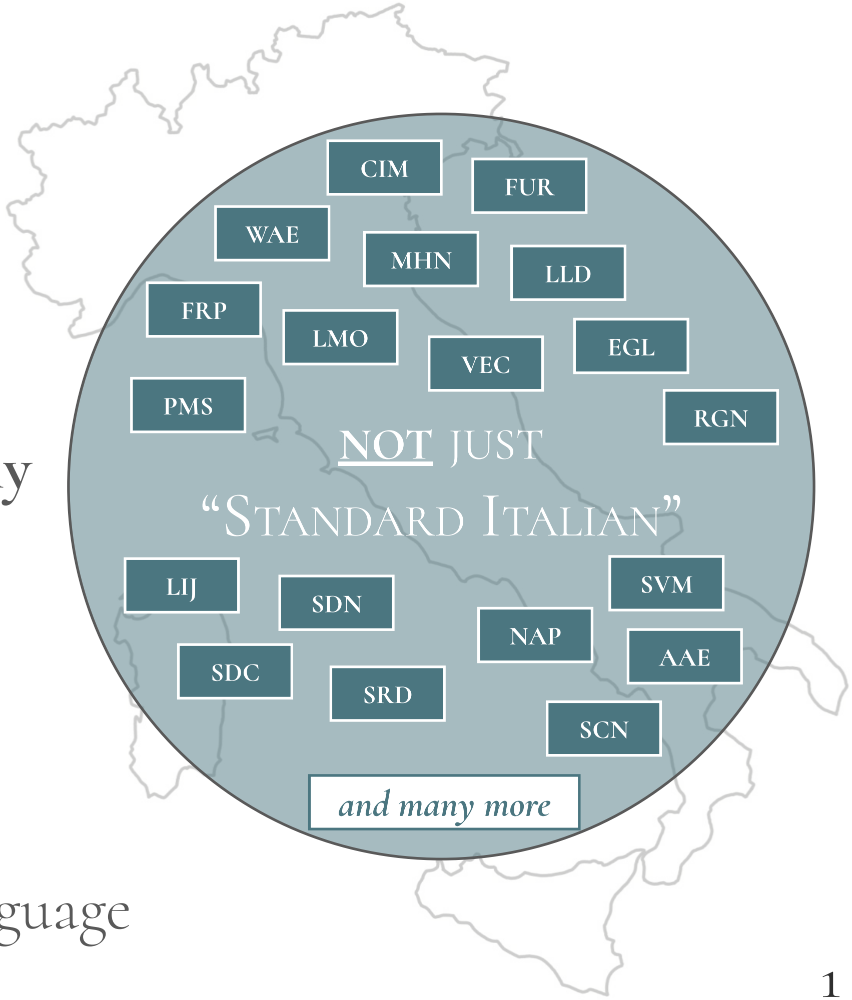
Introduction

Italy: linguistically-diverse country

- Many **languages**, **dialects**, and **regional varieties**
- Mostly **oral** and **without established orthography**

Diatopic language variation in Italy

- Focal point in **linguistics** (e.g., linguistic atlases)
- **User-generated texts**: informal, spontaneous language



Data

DiatopIt: the first social media corpus focused on **diatopic language variation in Italy** for *language varieties other than Standard Italian*

- Actual use, orthography choices, code-switching (*language contact* and *vitality*)

1 **chiov' tutt a jurnat', ce serv' o mbrell'**

en. *it's raining all day, we need an umbrella*

2 **ho così sonno che me bala l'oeucc**

en. *I'm so sleepy that my eye trembles*

3 **da caruso anche io ci andavo spesso!**

en. *I used to go there often as a kid too!*

Ramponi & Casula, 2023. "*DiatopIt: A Corpus of Social Media Posts for the Study of Diatopic Language Variation in Italy*". VarDial@EACL.

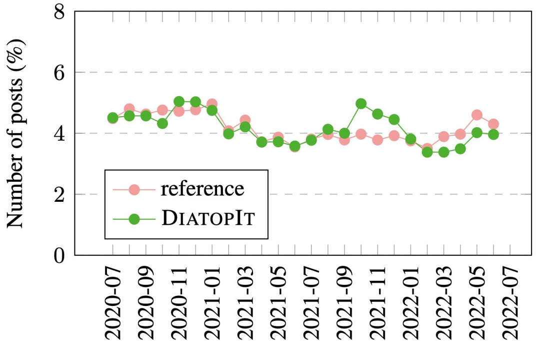
Data

Source: Twitter, geolocated in Italy
Timeframe: 2 years [2020-07 – 2022-06]
Sampling: based on (*curated*) OOV tokens
Curation: manual exclusion of spam/mismatches
Augmentation: for under-represented areas

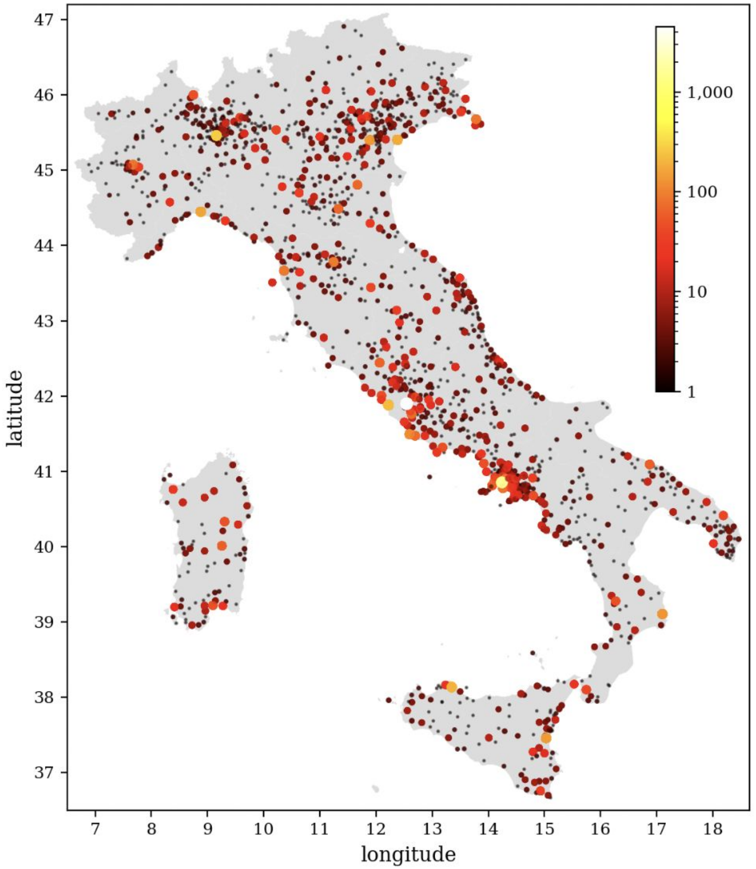
15K+ posts by 3,7K users

- *Thorough corpus analysis in Ramponi & Casula (2023)*

Distribution over time



Distribution over space



Ramponi & Casula, 2023. "DiatopIt: A Corpus of Social Media Posts for the Study of Diatopic Language Variation in Italy". VarDial@EACL.

Task description and evaluation

Given the text of a post exhibiting regional Italian features or (partially or fully) written in local languages and dialects of Italy, ***predict the location in which the variety expressed in the post is spoken***

Tracks: STANDARD TRACK (*country-level*) or SPECIAL TRACK (*linguistic area of choice*)

- **A) Coarse-grained geolocation**. Predict the region; macro F1 score
- **B) Fine-grained geolocation**. Predict the lat/lon coordinates; avg dist (km)

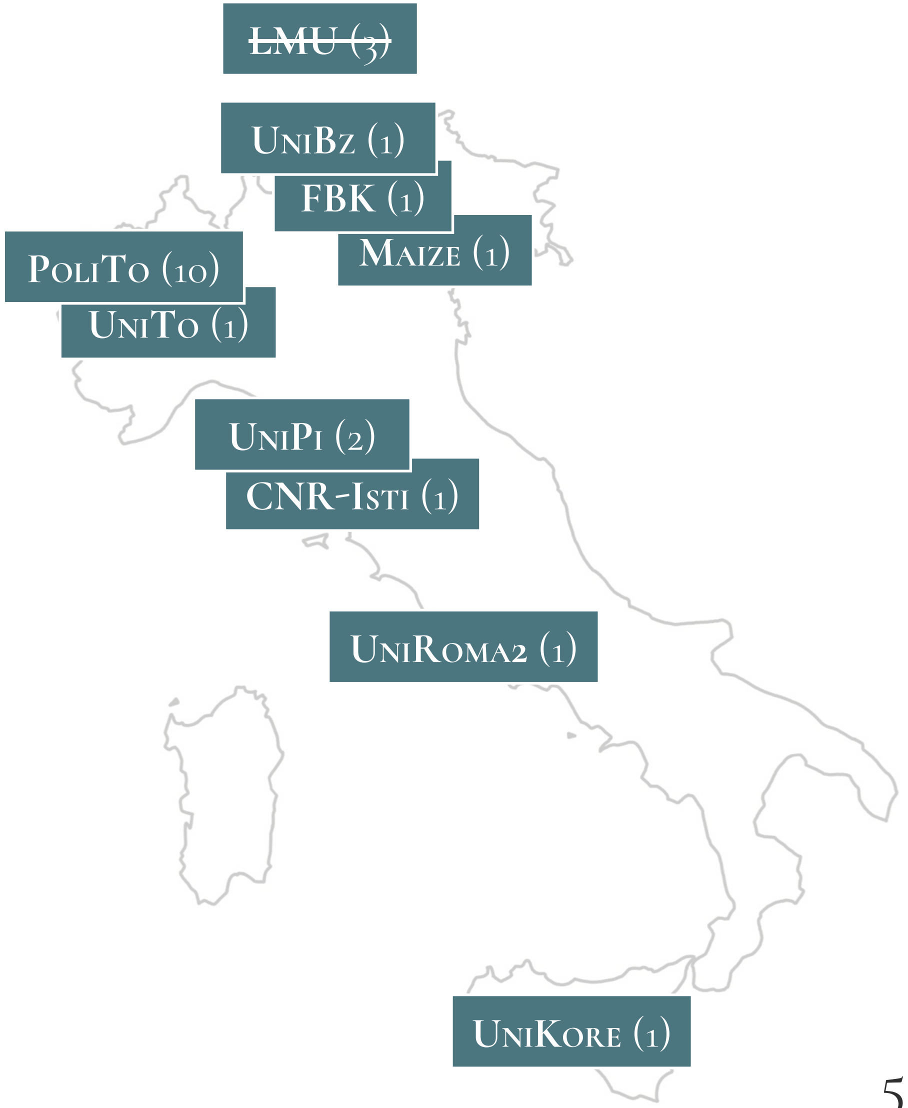
**Dev/test sets*: smoothed distribution, further curation to include a wide range of linguistic phenomena / microvariation*

Participation

37 registrations and **35 submitted runs**

Heterogeneously composed teams with up to 7 individuals, from master students to senior academic researchers

	Subtask A (*coarse-grained*)	Subtask B (*fine-grained*)	*Total*
Standard track	**14 (5)**	**12 (5)**	*26 (6)*
Special track	**6 (2)**	**3 (1)**	*9 (2)*
Total	*20 (5)*	*15 (5)*	*35 (6)*



Standard track

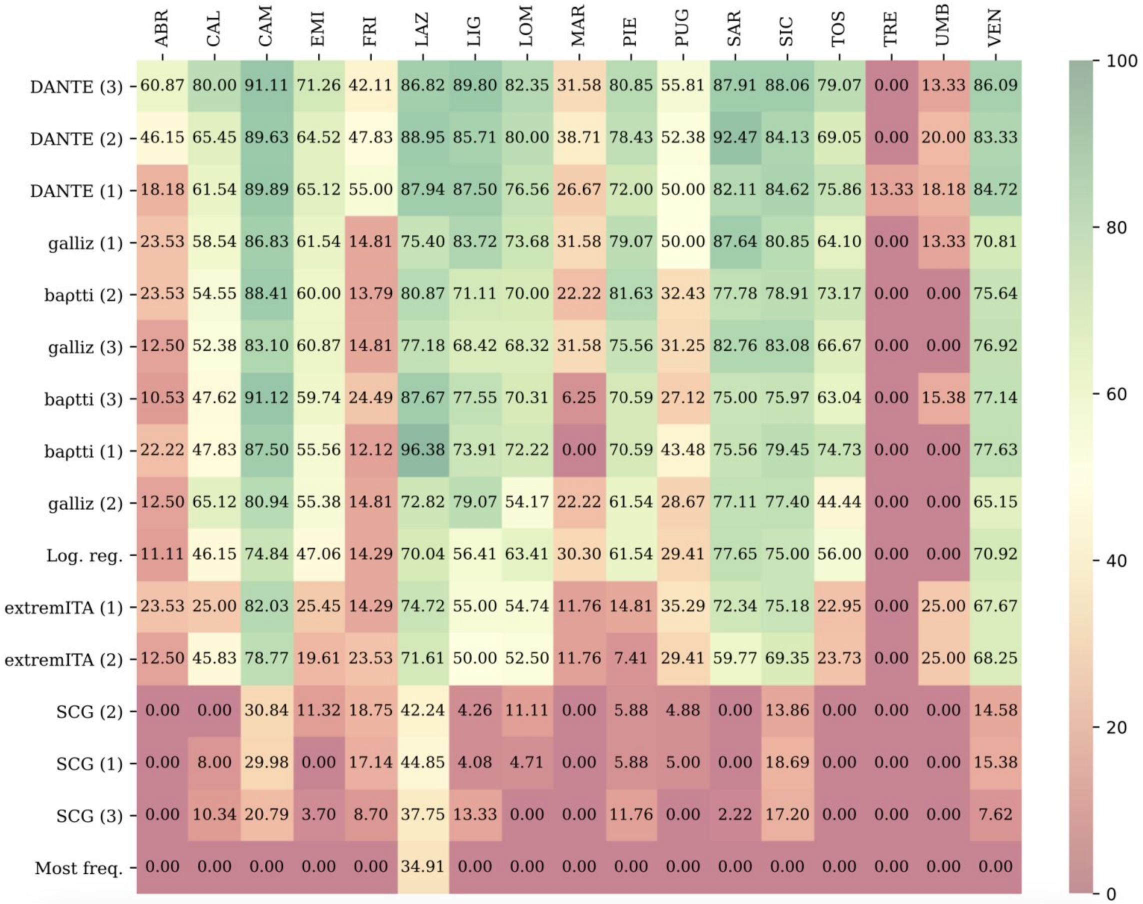
Subtask A: *Methods and results*

	Team	Run	P	R	F1	*Model & pretraining data*		*Extra pretraining methods & data*	*Fine-tuning methods and data*
1	DANTE	(3)	79.46	63.75	66.30	**Ens(15× 🙂)**	–	–	–
2	DANTE	(2)	66.98	62.65	63.93	🙂	it	MTL, sources: » W	STL, DiatopIt
3	DANTE	(1)	65.18	60.09	61.72	🙂	it	MTL, sources: » W	STL, DiatopIt
4	galliz	(1)	82.94	52.25	56.20	**Ens(🙂+📖)**	en	–	STL, *augmented* DiatopIt
5	baptti	(2)	67.97	51.62	53.18	🙂	it	CPT, *augmented* DiatopIt w/ » W	STL, DiatopIt
6	galliz	(3)	74.58	49.49	52.08	**Ens(🙂+📖)**	en	–	STL, *augmented* DiatopIt
7	baptti	(3)	52.93	51.75	51.74	**Ens(🙂+lr)**	it	–	STL, DiatopIt
8	baptti	(1)	56.05	51.68	51.72	🙂	it	–	MTL, DiatopIt
9	galliz	(2)	68.98	45.36	47.74	**Ens(🙂+📖)**	en	–	STL, *augmented* DiatopIt
	Log. reg.		*62.19*	*42.43*	*46.11*				
10	extremITA	(1)	72.14	38.84	39.99	T5	it	–	MTL, EVALITA 2023
11	extremITA	(2)	65.03	37.62	38.18	🦙	*	LoRA, Alpaca instructions in "it"	MTL, EVALITA 2023
12	SCG	(2)	12.92	9.82	9.28	svm	–	–	STL, DiatopIt
13	SCG	(1)	10.15	9.97	9.04	svm	–	–	STL, DiatopIt
14	SCG	(3)	10.42	6.60	7.85	lr	–	–	STL, DiatopIt
	Most freq.		*1.24*	*5.88*	*2.05*				

Models: **Ens(*)** Ensemble; 🙂 BERT-based; 🦙 LLaMa-based; 📖 dictionary-based (» + DiatopIt); T5 T5-based; svm Support vector machines; lr Logistic regression
Methods: **STL**: Single-task learning; **MTL**: Multi-task learning; **CPT**: Contrastive pretraining **Sources**: » Dialettando; W Wikipedia language editions

Standard track

Subtask A: *Analysis*



	ABR	CAL	CAM	EMI	FRI	LAZ	LIG	LOM	MAR	PIE	PUG	SAR	SIC	TOS	TRE	UMB	VEN
DANTE (3)	60.87	80.00	91.11	71.26	42.11	86.82	89.80	82.35	31.58	80.85	55.81	87.91	88.06	79.07	0.00	13.33	86.09
DANTE (2)	46.15	65.45	89.63	64.52	47.83	88.95	85.71	80.00	38.71	78.43	52.38	92.47	84.13	69.05	0.00	20.00	83.33
DANTE (1)	18.18	61.54	89.89	65.12	55.00	87.94	87.50	76.56	26.67	72.00	50.00	82.11	84.62	75.86	13.33	18.18	84.72
galliz (1)	23.53	58.54	86.83	61.54	14.81	75.40	83.72	73.68	31.58	79.07	50.00	87.64	80.85	64.10	0.00	13.33	70.81
baptti (2)	23.53	54.55	88.41	60.00	13.79	80.87	71.11	70.00	22.22	81.63	32.43	77.78	78.91	73.17	0.00	0.00	75.64
galliz (3)	12.50	52.38	83.10	60.87	14.81	77.18	68.42	68.32	31.58	75.56	31.25	82.76	83.08	66.67	0.00	0.00	76.92
baptti (3)	10.53	47.62	91.12	59.74	24.49	87.67	77.55	70.31	6.25	70.59	27.12	75.00	75.97	63.04	0.00	15.38	77.14
baptti (1)	22.22	47.83	87.50	55.56	12.12	96.38	73.91	72.22	0.00	70.59	43.48	75.56	79.45	74.73	0.00	0.00	77.63
galliz (2)	12.50	65.12	80.94	55.38	14.81	72.82	79.07	54.17	22.22	61.54	28.67	77.11	77.40	44.44	0.00	0.00	65.15
Log. reg.	11.11	46.15	74.84	47.06	14.29	70.04	56.41	63.41	30.30	61.54	29.41	77.65	75.00	56.00	0.00	0.00	70.92
extremITA (1)	23.53	25.00	82.03	25.45	14.29	74.72	55.00	54.74	11.76	14.81	35.29	72.34	75.18	22.95	0.00	25.00	67.67
extremITA (2)	12.50	45.83	78.77	19.61	23.53	71.61	50.00	52.50	11.76	7.41	29.41	59.77	69.35	23.73	0.00	25.00	68.25
SCG (2)	0.00	0.00	30.84	11.32	18.75	42.24	4.26	11.11	0.00	5.88	4.88	0.00	13.86	0.00	0.00	0.00	14.58
SCG (1)	0.00	8.00	29.98	0.00	17.14	44.85	4.08	4.71	0.00	5.88	5.00	0.00	18.69	0.00	0.00	0.00	15.38
SCG (3)	0.00	10.34	20.79	3.70	8.70	37.75	13.33	0.00	0.00	11.76	0.00	2.22	17.20	0.00	0.00	0.00	7.62
Most freq.	0.00	0.00	0.00	0.00	0.00	34.91	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Results differ a lot between regions

e.g., **Abr, Mar, Tre** and **Umb**
[scarce in `train`, absent in `dev`]
Very hard to model w/ traditional learning and tuning methods
– Beyond traditional learning/tuning

e.g., **Cal, Emi, Fri** and **Pug**
[represented in `train`, present in `dev`]
Easily misclassified w/ regions in which similar varieties are predominantly used
– Beyond “raw modeling”: linguistics!

Standard track

Subtask B: *Methods and results*

	Team	Run	Avg dist (km)	*Model & pretraining data*		*Extra pretraining methods & data*	*Fine-tuning methods and data*
1	baρtti	(3)	97.74	☺ + **Postproc**	it	–	**MTL**, DiatopIt
2	baρtti	(1)	98.79	☺	it	**CPT**, *augmented* DiatopIt w/ » W	**MTL**, DiatopIt
3	DANTE	(3)	110.35	**Ens**(2 × ☺)	it	–	–
4	DANTE	(2)	112.58	☺	it	**MTL**, sources: » W	**STL**, DiatopIt
5	DANTE	(1)	114.00	☺	it	**MTL**, sources: » W	**STL**, DiatopIt
6	baρtti	(2)	120.02	☺	it	–	**MTL**, DiatopIt
7	extremITA	(1)	126.10	T5	it	–	**MTL**, EVALITA 2023
8	Salogni	(1)	128.19	Ro☺a	–	–	**STL**, DiatopIt
9	extremITA	(2)	145.15	🦙	*	LoRA, Alpaca instructions in "it"	**MTL**, EVALITA 2023
	kNN		*263.35*				
10	SCG	(1)	280.99	lr	–	–	**STL**, DiatopIt
	Centroid		*281.04*				
11	SCG	(2)	281.20	lr	–	–	**STL**, DiatopIt
12	SCG	(3)	289.91	knn	–	–	**STL**, DiatopIt

Models: **Ens(*)** Ensemble; ☺ BERT-based; 🦙 LLaMa-based; **Postproc** Geographical postprocessing; T5 T5-based; knn *k*-nearest neighbors; lr Logistic regression
Methods: **STL**: Single-task learning; **MTL**: Multi-task learning; **CPT**: Contrastive pretraining **Sources**: » Dialettando; W Wikipedia language editions

Standard track

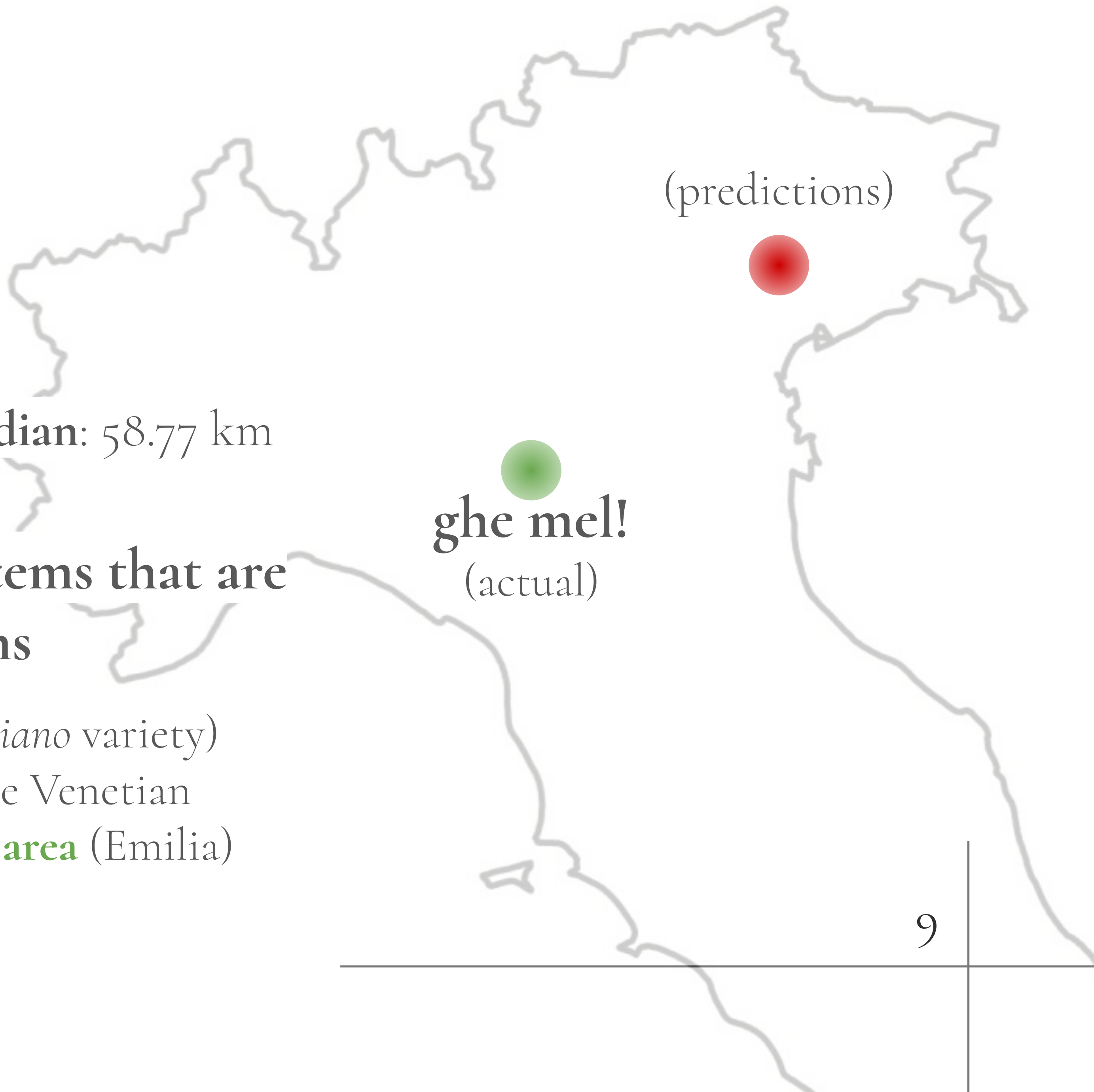
Subtask B: *Analysis*

Test set prediction (avg error)

- **Min**: 0.89 km, **max**: 668.11 km, **median**: 58.77 km

Misclassification due to **lexical items that are highly frequent in other locations**

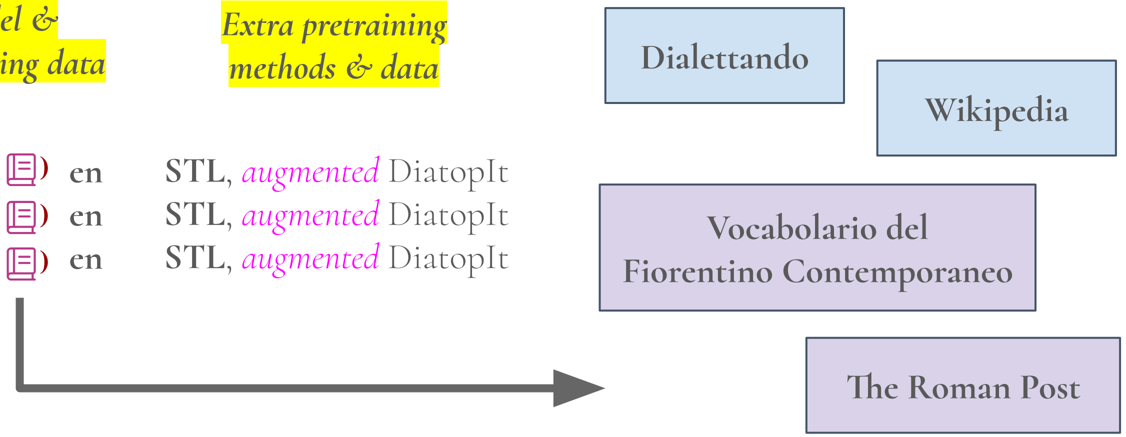
- e.g., "*ghe* mel!" (en: "of course", *Parmigiano* variety) in the **Treviso area** (Veneto, due to the Venetian ADV/PROP "ghe") instead of the **Parma area** (Emilia)



Special track

Subtask A and B: *Methods and results*

	Team	Run	P	R	F_1	*Model & pretraining data*		*Extra pretraining methods & data*
	TUSCANY-LAZIO AREA							
1	galliz	(3)	81.25	83.32	82.20	Ens(☺+🗎)	en	STL, *augmented* DiatopIt
2	galliz	(1)	72.43	80.42	73.40	Ens(☺+🗎)	en	STL, *augmented* DiatopIt
3	galliz	(2)	72.43	80.42	73.40	Ens(☺+🗎)	en	STL, *augmented* DiatopIt
	Log. reg.		*91.79*	*66.67*	*70.53*			
	Most freq.		*38.62*	*50.00*	*43.58*			



The **Gallo-Italic area** (PIE, LOM, LIG and EMI) has been explored, too

- **Subtask A**: SCG team, 3 runs based on Logistic Regression and Support Vector Machines
- **Subtask B**: SCG team, 3 runs based on Logistic Regression and k-Nearest Neighbors

Discussion and conclusion

- GeoLingIt has attracted **wide interest** from the community
- Modeling diatopic variation in Italy is a **difficult but exciting task**
- Great opportunities for more **linguistically-grounded NLP**

Poster booster session (~3 min each)

Baρtti team
A. Koudounas, F. Giobergia, I. Benedetto, S. Monaco, L. Cagliero, D. Apiletti, and E. Baralis

DANTE team
G. Gallipoli, M. La Quatra, D. Rege Cambrin, S. Greco, and L. Cagliero

Galliz team
T. Labruna and S. Gallo

Salogni team
I. Salogni