DiatopIt

A corpus of social media posts
for the study of diatopic language variation in Italy

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

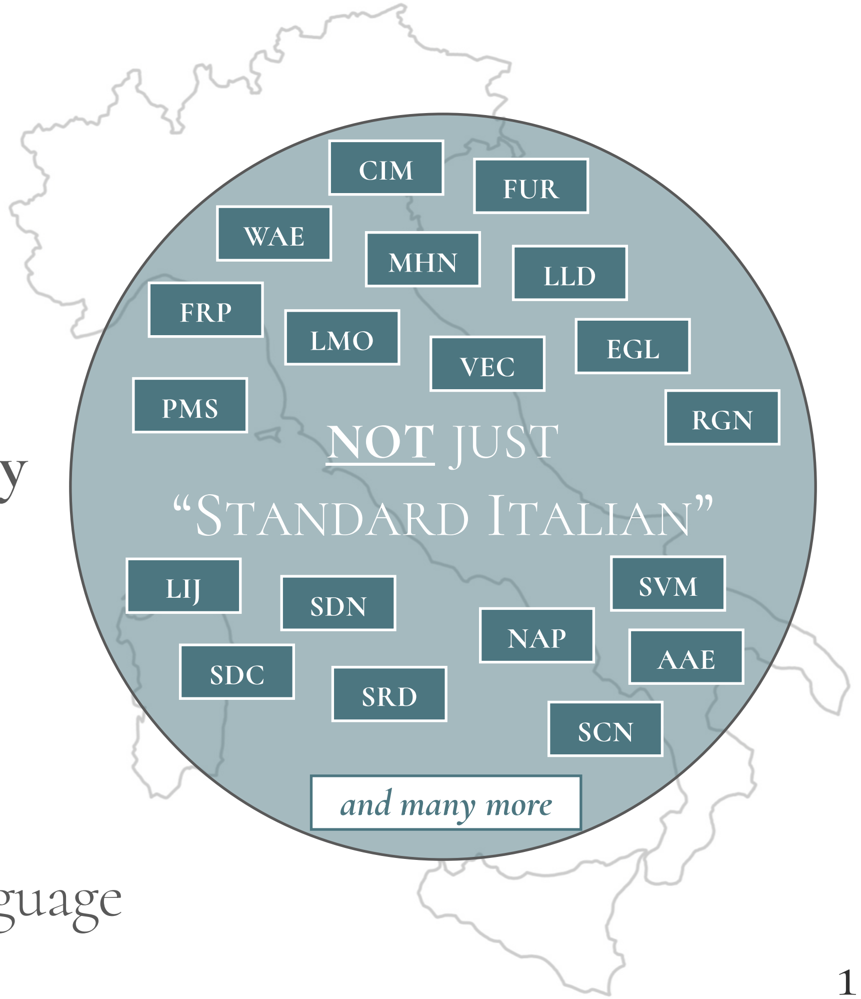
[1]Fondazione Bruno Kessler [2]University of Trento

VarDial 2023 – *Dubrovnik (HR)*

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

Contribution

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

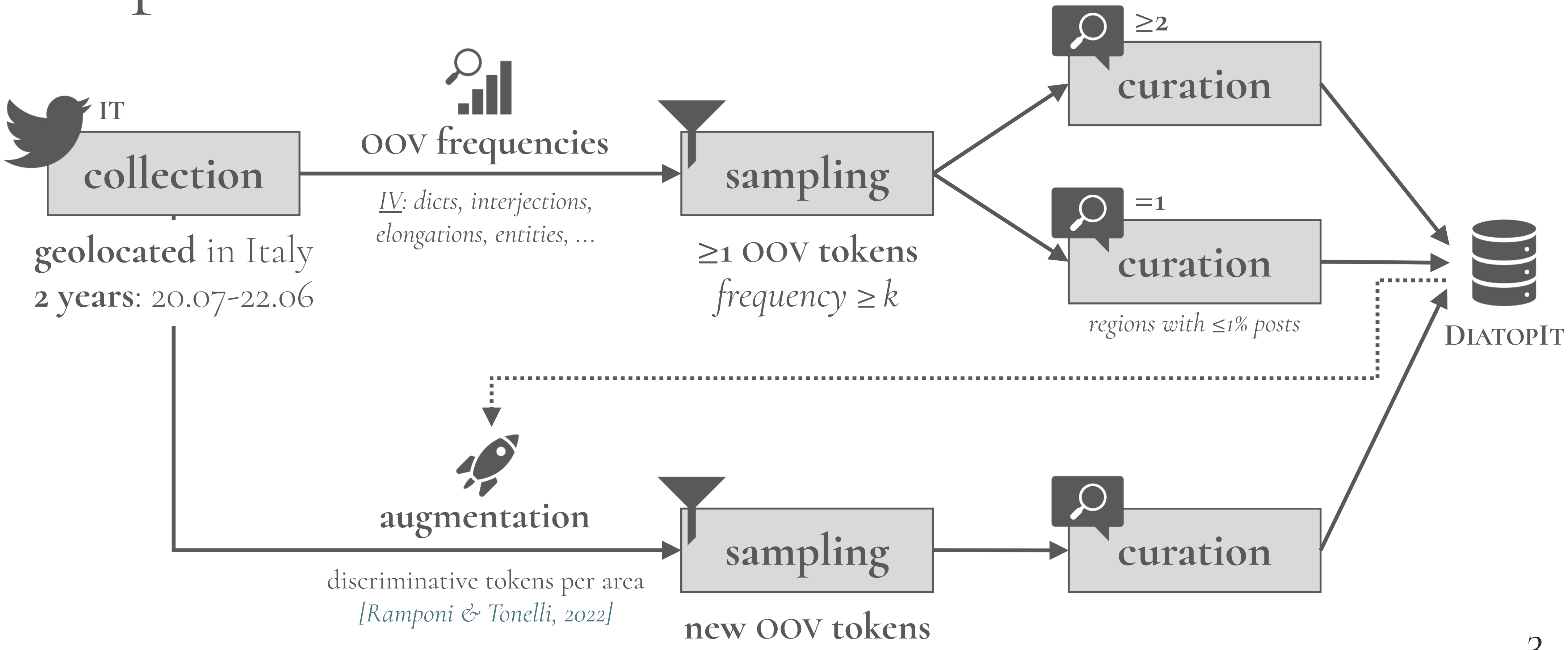
- Actual use, orthography choices, code-switching (*language contact* and *vitality*)

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!*

Corpus creation



IT

collection

geolocated in Italy

2 years: 20.07-22.06

OOV **frequencies**

IV: dicts, interjections, elongations, entities, …

sampling

≥1 OOV tokens

frequency ≥ k

EXCLUDE

posts by **spam** users

content-location **mismatches**

≥2

curation

=1

curation

regions with ≤1% posts

DIATOPIT

augmentation

discriminative tokens per area

[Ramponi & Tonelli, 2022]

sampling

new OOV tokens

curation

Ramponi & Tonelli, 2022. *"Features or Spurious Artifacts? Data-centric Baselines for Fair and Robust Hate Speech Detection". NAACL.*

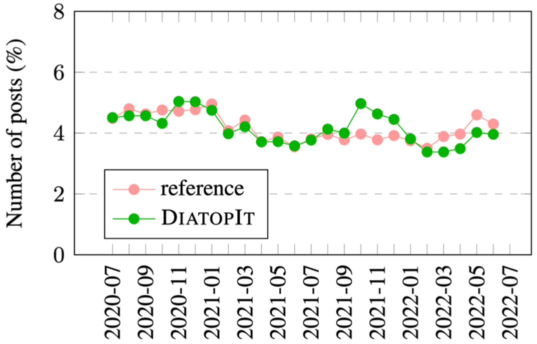
Corpus analysis: *distribution*

15K+ posts by 3,7K authors – *4.1 posts/user*

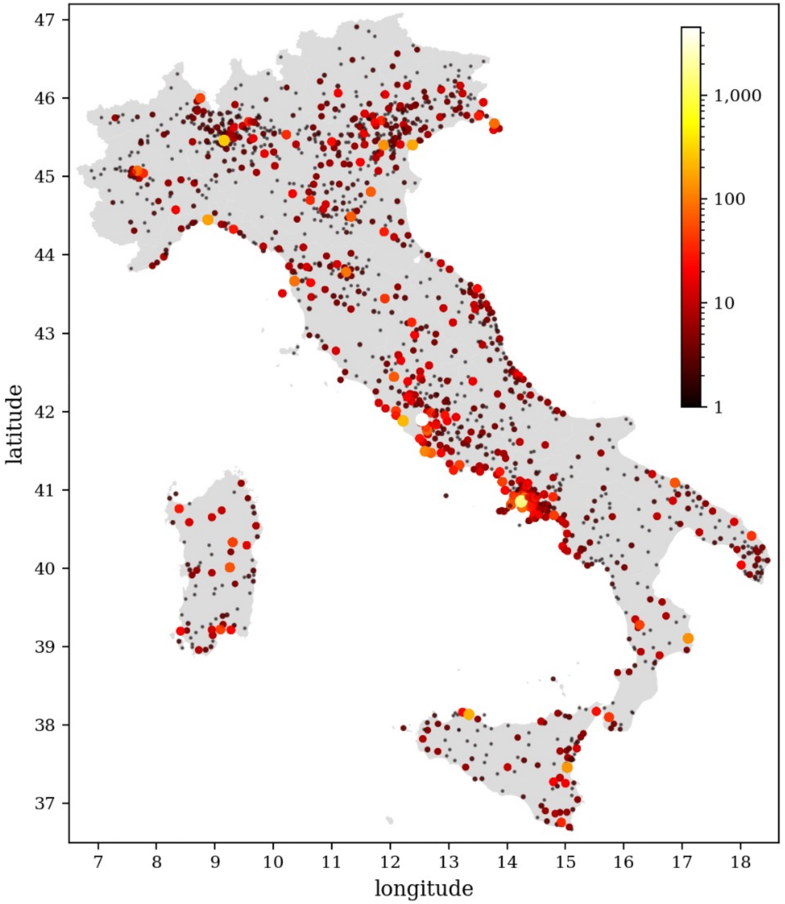
- **55K OOV tokens** – *14.1% avg OOV/post*
- *cities, coastal/lowlands* vs *rural/mountain* areas

Temporal biases minimized due to the topic-agnostic corpus creation procedure

Distribution over time



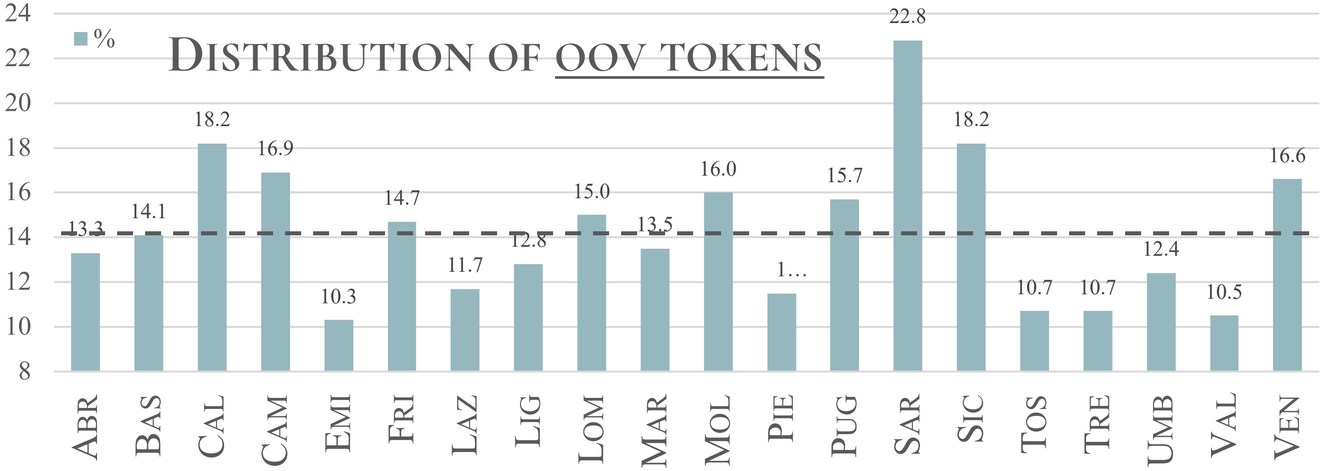
Distribution over space



Corpus analysis: *distribution*

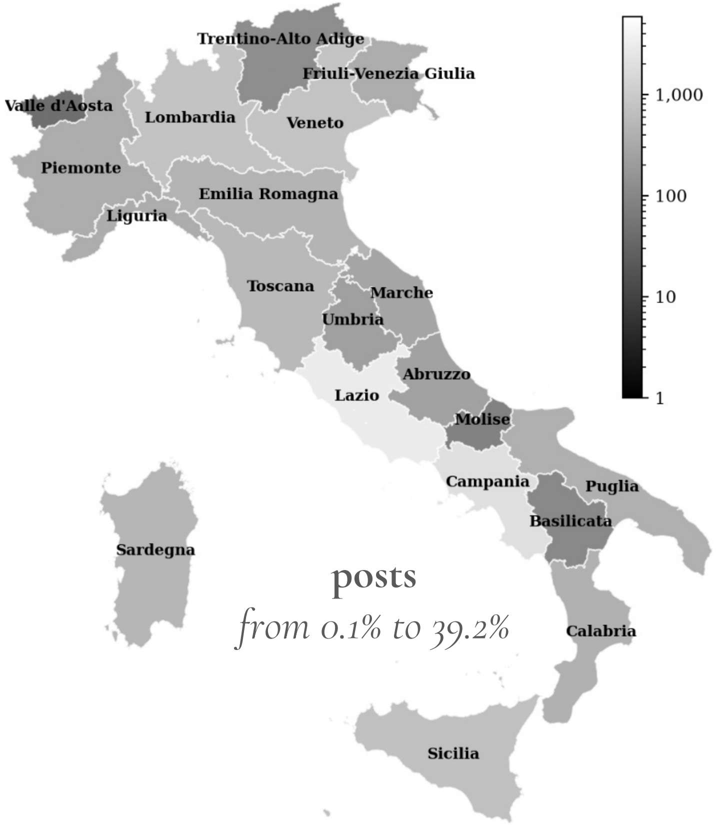
Regions varies a lot in terms of:

- **posts** – *population & use of varieties other than Italian*
- **OOV tokens** – *non-Standard Italian lexical items*



DISTRIBUTION OF OOV TOKENS

DISTRIBUTION OVER SPACE



posts
from 0.1% to 39.2%

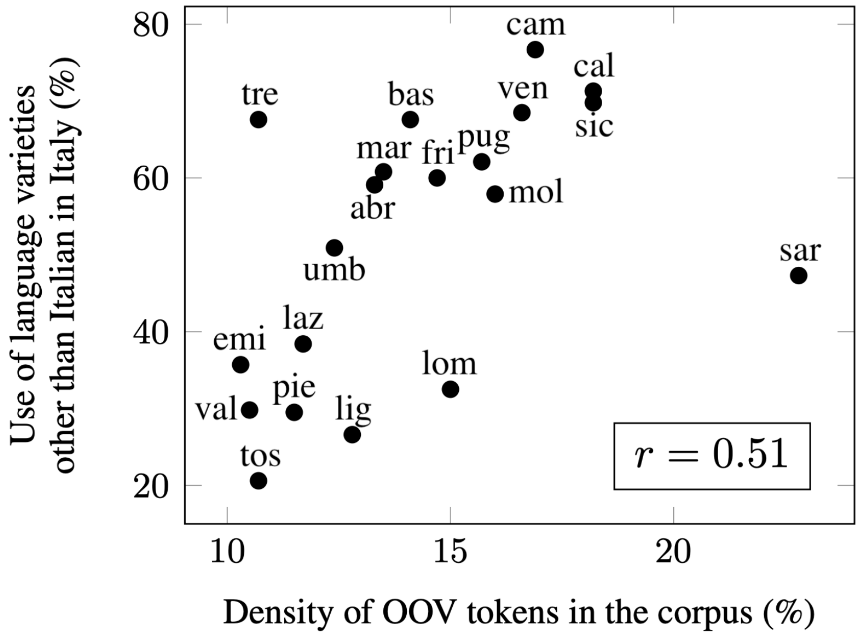
Corpus analysis: OOV *density and actual use*

Degree of mixing with Standard Italian

- HYP: the *more* a variety is used, the *less* lexical items that belong to Italian are employed

OOV **density** *vs* **use of varieties** [ISTAT, 2017]

- Substantial correlation (*r* = 0.51)
- SAR (*speakers' awareness*), TRE (*German varieties*)

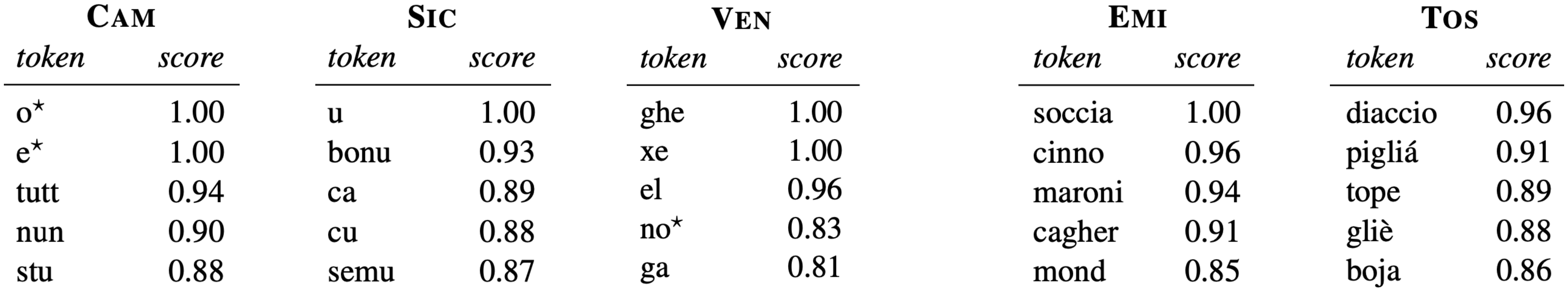


***ISTAT*, 2017**. "L'uso della lingua italiana, dei dialetti e di altre lingue in Italia". ISTAT website (accessed 2023-02-01).

Corpus analysis: *language use and vitality*

Most predictive per-region OOV tokens *[Ramponi & Tonelli, 2022]*

- HYP: the *more* language varieties are spoken in a region, the *higher* the likelihood that non-content OOV tokens (e.g., art, prep, conj) are used



CAM		**SIC**		**VEN**		**EMI**		**TOS**	
token	*score*	*token*	*score*	*token*	*score*	*token*	*score*	*token*	*score*
o*	1.00	u	1.00	ghe	1.00	soccia	1.00	diaccio	0.96
e*	1.00	bonu	0.93	xe	1.00	cinno	0.96	pigliá	0.91
tutt	0.94	ca	0.89	el	0.96	maroni	0.94	tope	0.89
nun	0.90	cu	0.88	no*	0.83	cagher	0.91	gliè	0.88
stu	0.88	semu	0.87	ga	0.81	mond	0.85	boja	0.86

confident use of local language varieties

restricted function of language varieties

Ramponi & Tonelli, 2022. "Features or Spurious Artifacts? Data-centric Baselines for Fair and Robust Hate Speech Detection". NAACL.

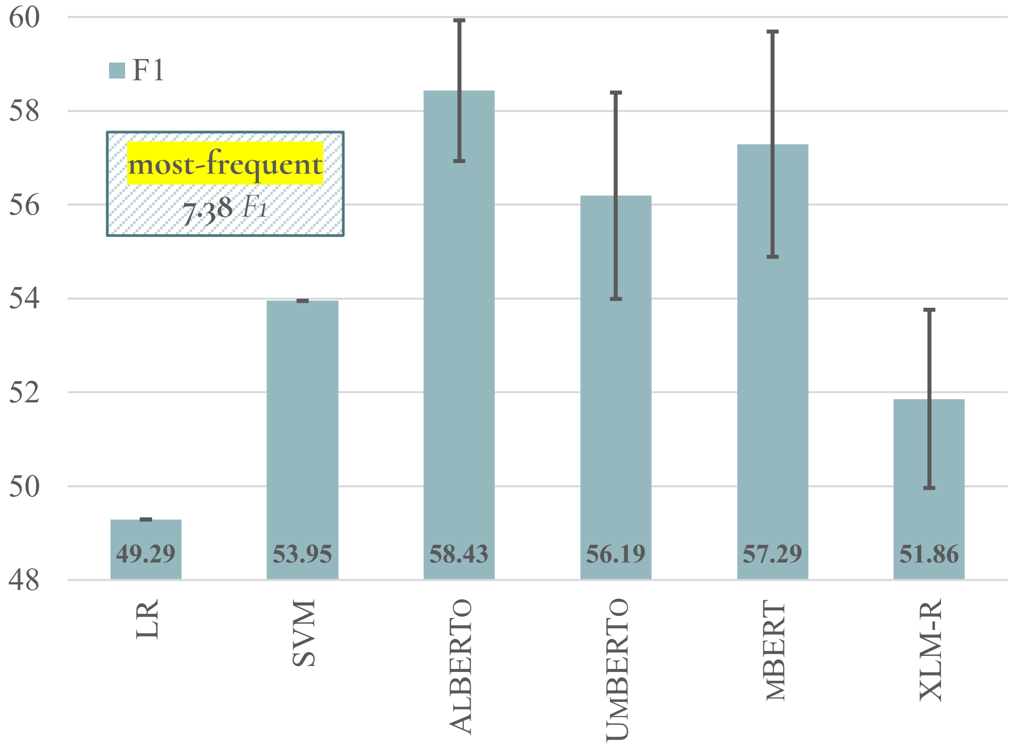
Experiments

How difficult is it to model diatopic language variation in Italy?

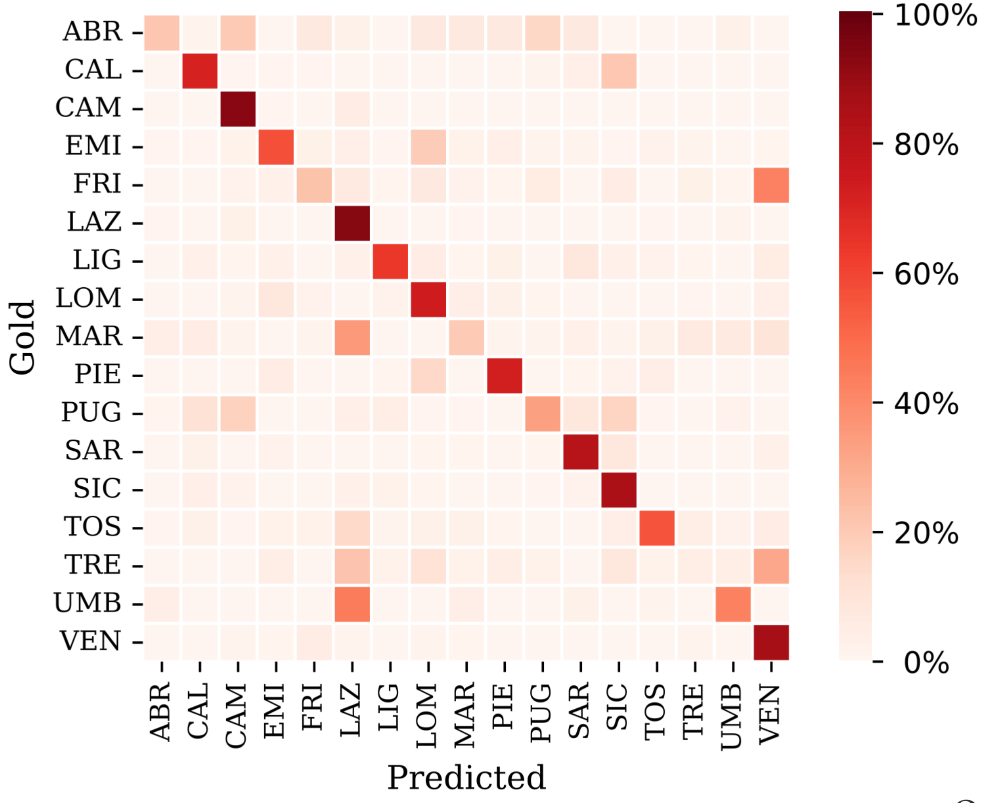
Experimental setup

- **Tasks**: *coarse-grained geolocation* [**CG**] and *fine-grained geolocation* [**FG**]
- **Evaluation**: *macro F1* [CG] and *avg dist (km)* [FG] *on regions with >50 total posts*
  - ***Dev/test* curation**: wide range of linguistic phenomena and microvariation
  - **Smoothed distribution**: more reliable evaluation for under-represented areas

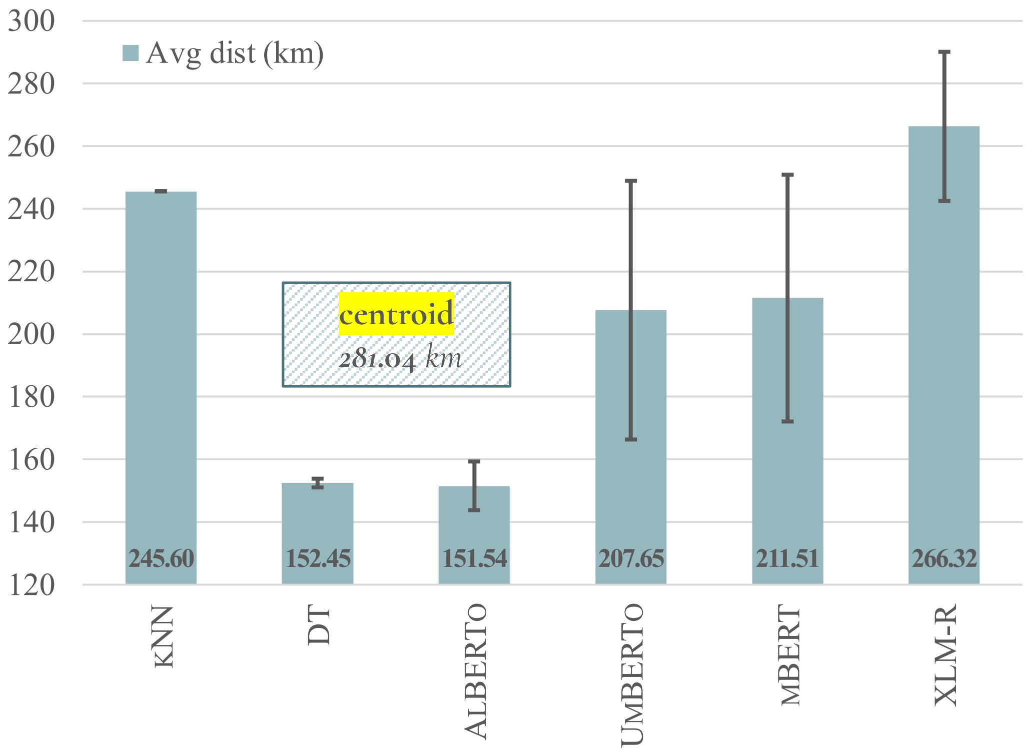
Results and discussion: *CG task*



most-frequent
7.38 *F1*



Results and discussion: *FG task*



Modeling diatopic variation is **hard** for transformer-based models, too

- Limited vocabulary coverage
- Pre-training data impacts stability

Conclusion

Corpus on *diatopic language variation* in Italy

- (*Partially and even fully*) written in local **languages**, **dialects**, and **regional Italian**
- **Spontaneous**: representative of actual use
- **Varied**: orthography and microvariation

NOT JUST
"STANDARD ITALIAN"

Useful to study diatopic variation, code-switching and divergences in orthography, in order to assess vitality across local language varieties