DIATOPIT: A Corpus of Social Media Posts for the Study of Diatopic Language Variation in Italy

Alan Ramponi,[☼] **Camilla Casula**[☼☂]
alramponi@fbk.eu, ccasula@fbk.eu

[☼] Fondazione Bruno Kessler (FBK), Italy
[☂] University of Trento, Italy

Abstract

We introduce DIATOPIT, the first corpus specifically focused on diatopic language variation in Italy for language varieties other than Standard Italian. DIATOPIT comprises over 15K geolocated social media posts from Twitter over a period of two years, including regional Italian usage and content fully written in local language varieties or exhibiting code-switching with Standard Italian. We detail how we tackled key challenges in creating such a resource, including the absence of orthography standards for most local language varieties and the lack of reliable language identification tools. We assess the representativeness of DIATOPIT across time and space, and show that the density of *non*-Standard Italian content across areas correlates with actual language use. We finally conduct computational experiments and find that modeling diatopic variation on highly multilingual areas such as Italy is a complex task even for recent language models.[1]

1 Introduction

Italy is one of the most linguistically-diverse countries in Europe. Despite its relatively small geographical area, it exhibits a notable profusion of linguistic variation, "*hold[ing] especial treasures for linguists*" (Maiden and Parry, 1997). Therefore, the study of diatopic linguistic variation in Italy has constantly been a focal point in linguistics (Bartoli et al., 1995; Jaberg et al., 1987, *inter alia*).

On the other hand, little attention has been given so far to this matter in the natural language processing community. Indeed, most work in NLP still focuses on Standard Italian (`ita`; the official national language), considering it as a "monolithic language". However, a large number of local languages, dialects, and regional varieties of Standard Italian (i.e., *regional Italian*)[2] shape the Italian linguistic landscape (Ramponi, 2022). Computational studies of diatopic variation can ultimately help to enrich and complement linguistic atlases, as well as to provide insights on actual use of local language varieties (e.g., adherence to orthography standards) and their vitality (e.g., code-switching as a sign of language replacement (Cerruti and Regis, 2005)). The ever-growing number of people who interact on social media offers opportunities in this direction, since user-generated texts are indeed informal, featuring linguistic patterns from spoken language (Eisenstein, 2013; van der Goot et al., 2021).

(a)	chiov' tutt a jurnat', ce serv' o mbrell'
	en. *it's raining all day, we need an umbrella*
(b)	ho così sonno che me bala l'oeucc
	en. *I'm so sleepy that my eye trembles*
(c)	da caruso anche io ci andavo spesso!
	en. *I used to go there often as a kid too!*

Table 1: Examples from DIATOPIT, with *non*-Standard Italian content in green. (a) posts fully written in local language varieties (here, Neapolitan [`nap`]); (b) posts code-switched with Standard Italian (here, Lombard [`lmo`]); (c) posts including regional Italian usage (here, "*caruso*" from the Sicilian [`scn`] "*carusu*"). Posts have been slightly redacted to preserve users' anonymity.

In this paper we introduce DIATOPIT, the first corpus of geolocated social media posts from Twitter with a focus on diatopic variation in Italy for language varieties[3] other than Standard Italian. DIATOPIT comprises 15,039 posts with content fully written in local language varieties (cf. Figure 1, *(a)*), exhibiting code-switching with Standard Italian (cf. Figure 1, *(b)*), or including regional Italian (cf. Figure 1, *(c)*). Compared to other existing datasets with geolocation information (Han et al., 2016; Gaman et al., 2020; Chakravarthi et al.,

[1] Repository: https://github.com/dhfbk/diatopit

[2] Geographical differentiation of Standard Italian due to influences by languages and dialects of Italy (Avolio, 2009).

[3] For brevity, we use "language varieties" to refer to local languages and dialects of Italy as well as regional Italian, whereas we add "local" to specifically refer to the former.

Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023), pages 187–199

2021), DIATOPIT is focused on Italy and on non-standard language use. We describe how we tackled challenges in the corpus creation process, such as the lack of reliable, variation-informed language identification tools and the absence of orthography standards for most local varieties (Section 2), and provide detailed analyses over time and space, also highlighting the density and function of *non*-Standard Italian content across Italian regions (Section 3). Finally, we show that modeling diatopic language variation is a difficult task even for state-of-the-art language models (Section 4).

The corpus is meant to encourage research on diatopic variation in Italy, study code-switching and divergences in orthography for local language varieties, and serve as a basis for responsible development of annotated resources for Italy's varieties.

2 Corpus Creation

Building a corpus of social media posts written in language varieties of Italy other than Standard Italian is a tough task, especially in the absence of reliable language identification tools.[4] Most languages and dialects of Italy – see Ramponi (2022) for an overview – are primarily oral and have no established orthography, and standards that have been proposed for a fraction of them are rarely adopted by their speakers. Indeed, when those language varieties are transposed into writing, speakers typically write "the way words sound" (Ramponi, 2022). The language functions of those varieties – most of which are endangered (Moseley, 2010) – are increasingly restricted, resulting in frequent code-switching with Standard Italian, a sign of language replacement (Cerruti and Regis, 2005).

In this section we describe how we tackle these challenges to build the DIATOPIT corpus. We detail all stages, from data collection (Section 2.1) and sampling for *non*-Standard Italian content (Section 2.2), to content curation and data augmentation of under-represented speaking areas (Section 2.3). Data statements (Bender and Friedman, 2018) for DIATOPIT are presented in Appendix A.

2.1 Collection of Geolocated Posts in Italy

For our initial collection, we use the Twitter APIs to retrieve geolocated tweets in Italy over a period of two years, from `2020-07-01` to `2022-06-30`. This ensures that coordinates of tweets fall within the Italian territory, and thus that content exhibiting linguistic variation is relevant to Italy. Moreover, the large time frame mitigates potential biases in the corpus about exceptional or occasional events, whereas the presence of the same number of months across years avoids over-representing recurring events, both local (e.g., the *Italian Song Festival*, February) and global (e.g., Christmas).

We then sample posts that have been classified as "`it`" by Twitter, due to the frequent code-switching of local language varieties with Standard Italian (cf. Section 2) and the absence of dedicated language classifiers. In addition, we observed that content (partially and even fully) written in language varieties of Italy is typically classified as `it` by the Twitter language identifier.[5] We obtain over 10 million geolocated tweets for further filtering.

2.2 Sampling *Non*-Standard Italian Posts

To construct a representative sample of social media posts written in language varieties of Italy other than Italian, we take our initial collection (Section 2.1) and further filter it to contain *non*-Standard Italian content. We deliberately avoid using predefined lexicons for sampling, since *(i)* their coverage is typically low in terms of both vocabulary and representation of local variants, and *(ii)* using them for sampling could bias our corpus towards standard orthographies, thus excluding variation due to speakers' lack of knowledge of written conventions (if any). We instead adopt a complementary approach in which lexical units for sampling naturally emerge from their actual use on social media.

We analyze the whole collection of tweets, computing frequencies of out-of-vocabulary (OOV) tokens.[6] We consider a token as OOV if it is not a special token (i.e., hashtag, punctuation, number, emoji) nor is part of the Aspell dictionary for Italian.[7] Additionally, we do not consider as OOV all tokens that are part of the English dictionary[8] to avoid including international discourse in our corpus. We inspect the resulting token frequencies and further exclude common interjections (e.g., *boh*, en: *I don't know*), elongated words (e.g., *ciaoo*, en: *helloo*), words in Italian with wrong diacritics

[4] Language identification tools for (a subset of) language varieties of Italy are mostly trained on Wikipedia, a very specific domain that does not reflect how those languages and dialects are typically used by their speakers (Ramponi, 2022).

[5] Nonetheless, in future work we plan to extend the corpus with the fraction of relevant content classified as non-`it`, too.

[6] Tokenization of posts has been performed by using the `it_core_news_sm` model by spaCy (`https://spacy.io`).

[7] `http://aspell.net`: `aspell6-it-2.2_20050523-0`.

[8] `http://aspell.net`: `aspell6-en-2020.12.07-0`.

(e.g., *perchè*; en: *why/because*), youth language and slang words (e.g., *xke*, en: *why/because* [ABBR.]; *buongiornissimo*, en: *good morning* [SUP.]), tokenization errors (e.g., *∼il*, en: *∼the*), tokens in foreign languages (e.g., *gracias*, en: *thank you*), tokens in Italian or English that are not included in Aspell dictionaries (e.g., *quest'*, en: *this* [CONTR.]; *t-shirt*), and tokens that explicitly refer to named entities (e.g., soccer players, singers, brands, cities).

We use tokens $t \in T_{oov}$ with a frequency $F(t) \geq k$ as our search keywords, and retain from the collection all tweets that contain at least one of such terms. To avoid including social media posts with tokenization errors and rare typos, we empirically set $k = 48$, which corresponds to an average token frequency of 2 occurrences per month. We obtain over 100K tweets with 953 search tokens. Search tokens are made available in our repository.

2.3 Corpus Curation and Augmentation

Posts that match at least one OOV token do not necessarily contain lexical items of local language varieties or signal of interest for diatopic studies. Indeed, our initial exploration revealed that a fraction of matched posts were spam messages or still contained no signal due to the ambiguity of some search terms. Moreover, we found occasional mismatches between the geolocation attached to posts and the language varieties used within them.[9]

Motivated by these factors, we focus on the subset of posts matching at least 2 OOV tokens (i.e., roughly 20K tweets) and conduct a manual curation process. Two curators with good knowledge of language varieties of Italy and background in NLP and sociolinguistics identified all user IDs whose posts contain *(i)* spam content or *(ii)* content in language varieties that are not spoken in the area of the geolocated position (e.g., due to tourism or relocation). We then removed all the tweets posted by spam users, the subset of posts with clearly incongruous content and geolocation, as well as matched tweets exhibiting no diatopic signals.

To mitigate the under-representation in our corpus of some areas in which local language varieties are scarcely spoken, we additionally conducted two steps of data augmentation. In the first step, the curators manually checked the remaining subset of posts with just a single matched OOV token for all regions with $\leq 1\%$ posts over the total.[10] During the whole process, cases of doubt were managed by the curators by consulting dictionaries and asking native speakers for clarification. Posts containing content in *non*-Standard Italian were then added to the corpus. In the second step, we took the set of tweets from all regions except the over-represented ones (i.e., Lazio and Campania; cf. Figure 2a) and employed the lexical artifacts package (Ramponi and Tonelli, 2022) to compute a ranking of the highly-discriminative tokens for each region in a *one-vs-rest* scheme. A list comprising the top 50 OOV tokens of each region, totalling 820 unique keywords, was then used to sample additional tweets from the initial collection (Section 2.1). The curators then manually checked these sets, adding relevant tweets to the corpus. Finally, we deduplicated the corpus by removing tweets that had the same content and author ID.[11]

3 Corpus Analysis

In this section we present detailed analyses on the DIATOPIT corpus. We first provide summary statistics (Section 3.1). Then, we discuss the corpus distribution across time and space (Section 3.2). Lastly, we show that the density of *non*-Standard Italian tokens across regions correlates with the actual use of languages varieties in Italy, and that language functions of the most indicative tokens per region are good indicators of vitality (Section 3.3).

3.1 Summary Statistics

In Table 2 we present summary statistics and density information about the corpus. DIATOPIT comprises 15,039 posts with geolocation information across all 20 administrative regions of Italy, accounting for a total of 388,069 tokens, 54,635 of which are OOV (i.e., 14.1%). Posts have an average length of 25.8 tokens and have been written by 3,672 authors (i.e., 4.1 posts per author on average).

By a closer look, Lazio (LAZ) and Campania (CAM) are the most represented regions in the corpus, with 39.2% and 21.5% instances, respectively. All other regions comprise from 0.1% to 5.9% posts, with those with $\leq 1.5\%$ instances representing territories with a small population or in which local language varieties are little spoken.

[9] Although language and mobility is an interesting topic, it goes beyond the purpose of this work. We leave the study of this phenomenon as future direction for research.

[10] We refer the reader to Appendix B for additional details.

[11] Indeed, we do not consider a tweet with the same content but posted by different authors as a duplicate, but rather a useful signal for diatopic studies and language vitality assessments, especially if posted from different locations.

	Instances		Tokens				Authors	Density		
	I (#)	I (%)	T_{all}	T_{all}^{unique}	T_{oov}	T_{oov}^{unique}	A	T_{all}/I	I/A	$\frac{T_{oov}}{T_{all}}$ (%)
ABR	166	1.1%	3,939	1,495	523	370	86	23.7	1.9	13.3%
BAS	49	0.3%	1,166	575	164	141	30	23.8	1.6	14.1%
CAL	336	2.2%	7,683	2,626	1,399	872	101	22.9	3.3	18.2%
CAM	3,240	21.5%	78,233	11,627	13,185	3,889	645	24.2	5.0	16.9%
EMI	395	2.6%	9,861	2,902	1,020	589	173	25.0	2.3	10.3%
FRI	270	1.8%	6,851	2,360	1,008	652	83	25.4	3.3	14.7%
LAZ	5,895	39.2%	162,532	19,379	19,031	4,635	987	27.6	6.0	11.7%
LIG	273	1.8%	6,378	1,853	819	434	82	23.4	3.3	12.8%
LOM	803	5.3%	20,966	5,125	3,139	1,535	327	26.1	2.5	15.0%
MAR	197	1.3%	5,035	1,821	679	432	96	25.6	2.1	13.5%
MOL	35	0.2%	692	364	111	90	21	19.8	1.7	16.0%
PIE	288	1.9%	6,498	2,094	750	434	127	22.6	2.3	11.5%
PUG	320	2.1%	8,000	2,558	1,254	733	157	25.0	2.0	15.7%
SAR	440	2.9%	11,711	3,513	2,665	1,504	129	26.6	3.4	22.8%
SIC	720	4.8%	16,780	4,355	3,050	1,444	240	23.3	3.0	18.2%
TOS	506	3.4%	13,640	3,449	1,459	700	194	27.0	2.6	10.7%
TRE	61	0.4%	1,434	670	153	111	37	23.5	1.6	10.7%
UMB	150	1.0%	4,129	1,425	512	284	49	27.5	3.1	12.4%
VAL	14	0.1%	420	260	44	42	14	30.0	1.0	10.5%
VEN	881	5.9%	22,121	5,093	3,670	1,593	252	25.1	3.5	16.6%
ALL	15,039	100.0%	388,069	40,744	54,635	16,482	3,672	25.8	4.1	14.1%

Table 2: Summary statistics for the DIATOPIT corpus. Region names (*left*) are presented with their first three letters (see Figure 2a for full names and location). Columns (*top*). I: instances (#: raw number; %: percentage); T_{all}: tokens; T_{all}^{unique}: unique tokens; T_{oov}: OOV tokens; T_{oov}^{unique}: unique OOV tokens; A: authors; T_{all}/I: average tokens per instance; I/A: average instances per author; T_{oov}/T_{all} (%): average density of OOV tokens within posts.

Regions vary a lot in terms of average density of OOV tokens within posts (T_{oov}/T_{all}). Sardinia (SAR), Sicilia (SIC), Calabria (CAL), Campania (CAM) and Veneto (VEN) are the regions in which lexical items of language varieties of Italy other than Standard Italian are used more frequently.[12] Lastly, LAZ, CAM, and VEN are the regions in which the ratio of instances per author (I/A) is higher, a sign of a more confident use of local language varieties by their speakers.

3.2 Distribution Across Time and Space

In order to assess the potential presence of temporal biases in our corpus, we examine the distribution of social media posts across time, and compare it with that of the initial collection (cf. Section 2.1). Figure 1 shows the percentage of tweets for each month within the 2-year time span for the DIATOPIT corpus and the reference (i.e., the initial collection). We observe that the number of posts in DIATOPIT closely follows the distribution of the reference, with the only exception for the period from `2021-10` to `2021-12`. We examined tweets posted within this time span and we positively found that the small peak is due to some users posting more than average about a wide range of topics rather than due to period-specific biases.

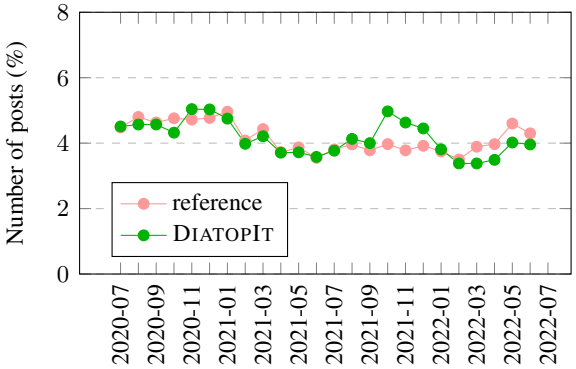
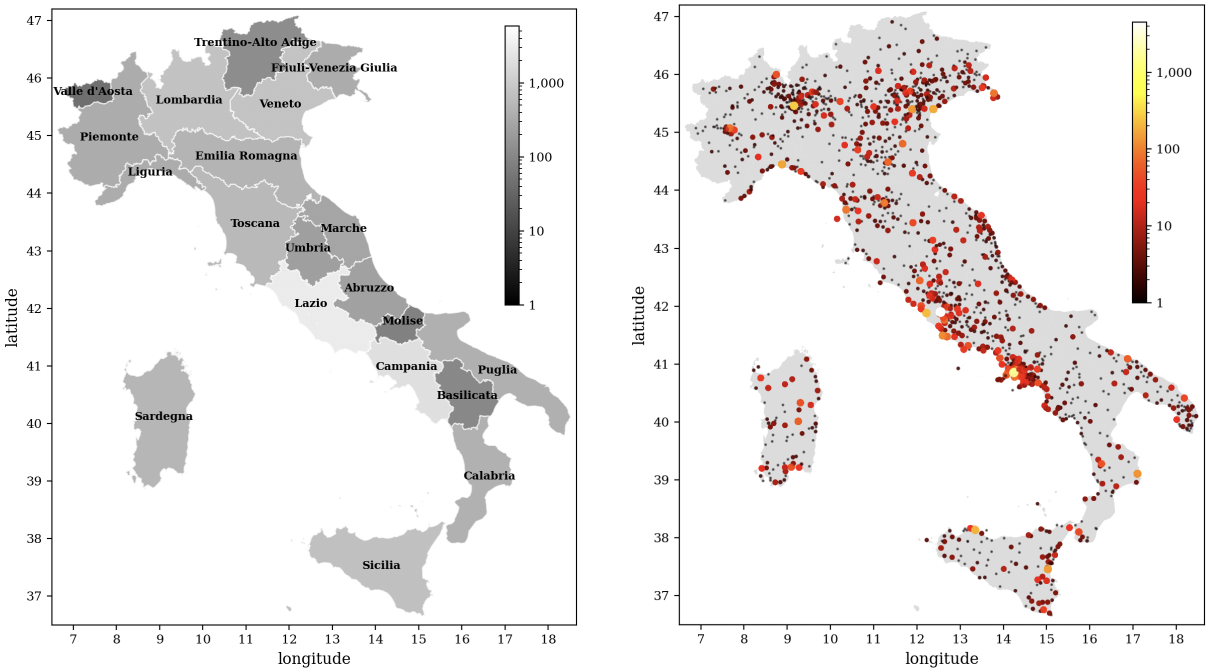


Figure 1: Distribution of social media posts over time in both DIATOPIT and the initial collection (*reference*).

As regards the spatial dimension, in Figure 2

[12] Note that multiple local languages and dialects are often spoken within a region, and they often cross administrative borders. Refer to Pellegrini (1977) for a linguistic map.



(a) Tweets by administrative region.

(b) Tweets by geographical coordinates.

Figure 2: Distribution of social media posts in the DIATOPIT corpus by administrative region and coordinates.

we present the distribution of tweets in our corpus. While Figure 2a contextualizes across space the per-region instances presented in Table 2, Figure 2b shows a fine-grained distribution of social media posts by geographical coordinates. As expected, a large number of posts comes from densely-populated cities and coastal and lowlands areas. Rural and mountain areas are instead weakly represented. Although the resident population is a good indicator for the amount of content that is posted online within a particular area, the density of *non*-Standard Italian content can diverge a lot between regions (cf. Section 3.1). Moreover, densely-populated areas do not always exhibit a high proportion of tweets. This is the case of e.g., Piemonte (PIE), a region of northwest Italy (cf. Figure 2a) with a population of > 4.2M, for which there exists a relatively low number of tweets containing *non*-Standard Italian content (1.9%, cf. Table 2) due to the limited use of local varieties (Figure 3).

3.3 Density and Functions of OOV Tokens

We hypothesize that geographical areas in which local language varieties are spoken the most are likely to exhibit a lower degree of mixing with Standard Italian compared to areas in which those are gradually disappearing. Indeed, the less a variety is used, the more lexical items that belong to Standard Italian would be employed.

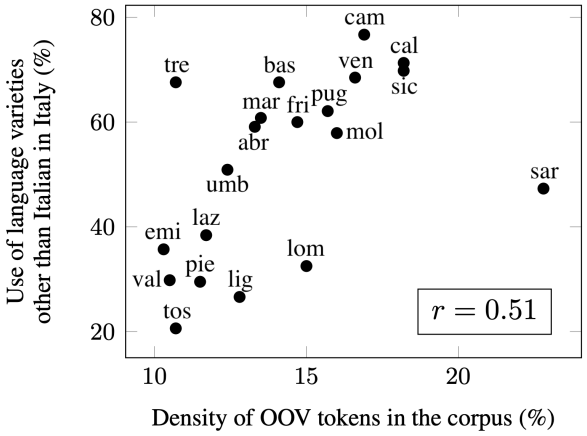


Figure 3: Pearson correlation coefficient (r) between the density of OOV tokens (T_{oov}/T_{all}) for each region and the actual usage of language varieties other than Standard Italian in those regions (ISTAT, 2017).

To test this hypothesis and assess the representativeness of DIATOPIT, we take the results of the most recent national survey on the actual use of languages and dialects in Italy divided by region (ISTAT, 2017) and check if the proportion of OOV tokens (T_{oov}/T_{all}) in our corpus for those regions correlates with it (cf. Appendix C). We calculated the Pearson correlation coefficient r and found a substantial correlation ($r = 0.51$). As shown in Figure 3, there is a high correlation for most regions, with the exception of Trentino-Alto Adige (TRE)

CAL		CAM		EMI		LAZ		LOM	
token	*score*	*token*	*score*	*token*	*score*	*token*	*score*	*token*	*score*
u	1.00	o⋆	1.00	soccia	1.00	na	1.00	i⋆	1.00
ccu	0.91	e⋆	1.00	cinno	0.96	de	0.97	el	0.99
i⋆	0.90	tutt	0.94	maroni	0.94	pe	0.97	ratt	0.99
frica	0.85	nun	0.90	cagher	0.91	je	0.94	ciapa	0.96
ca	0.84	stu	0.88	mond	0.85	er	0.89	inscì	0.93
PUG		**SAR**		**SIC**		**TOS**		**VEN**	
token	*score*	*token*	*score*	*token*	*score*	*token*	*score*	*token*	*score*
lu	1.00	su⋆	1.00	u	1.00	diaccio	0.96	ghe	1.00
sule	0.83	sa	0.99	bonu	0.93	pigliá	0.91	xe	1.00
ientu	0.82	tottu	0.97	ca	0.89	tope	0.89	el	0.96
trmon	0.74	itte	0.95	cu	0.88	gliè	0.88	no⋆	0.83
trimone	0.72	unu	0.93	semu	0.87	boja	0.86	ga	0.81

Table 3: Top-5 most indicative tokens and associated scores (in $[0, 1]$) for regions with $\geq 2\%$ instances in the DIATOPIT corpus. Tokens marked with ⋆ are those that are typically included in stopword lists for Standard Italian.

and Sardegna (SAR). While results for TRE can be justified by highly-spoken German varieties in the South Tyrol province that are little represented in our corpus (cf. Limitations section), we argue that results for SAR are due to the long-established speakers' awareness of the prestige status of their varieties (i.e., Sardinian: `srd`, Sassarese: `sdc`, and Gallurese: `sdn`). Indeed, the survey by ISTAT (2017) mostly framed questions using the word "*dialect*", a term that historically carries negative connotations in Italy (Avolio, 2009).

Besides the raw density of *non*-Standard Italian content, the function of the most indicative OOV tokens for each region can give insights into language use and vitality, too. Intuitively, the more language varieties are spoken in a region, the higher is the likelihood that non-content tokens that are necessary to form articulated sentences (e.g., articles, prepositions and conjunctions) are used.

To the goal, we employ the lexical artifacts package (Ramponi and Tonelli, 2022) and compute the most discriminative tokens for each region in a *one-vs-rest* scheme, i.e., unveiling lexical items that are more frequently used in the region of interest compared to all other regions. We present the top-5 most indicative tokens for all regions with $\geq 2\%$ instances over the total[13] in Table 3.

Regions in which local varieties are spoken the most (i.e., CAL, CAM, SIC, VEN; cf. Figure 3, *top*) mostly present non-content tokens as the most informative, confirming our hyphothesis. Both CAL and SIC have "*u*" (en: *the* [M. SG.]) and "*ccu/cu*" (en: *with*) among the most indicative terms, as well as "*i*" (CAL; en: *the* [M. PL.]), and "*semu*" (SIC; en: *we are*), amongst others. Relevant examples for CAM and VEN also include "o" and "el" (en: *the* [M. SG.]), "stu" (CAM; en: *this*), "ghe" (VEN; en: *there is*), and "xe" (VEN; en: *is*). SAR also shows non-content tokens as the most informative, e.g., "*su*" (en: *the* [M. SG.]), "*sa*" (en: *the* [F. SG.]) and "*unu*" (en: *a/an/one*), confirming that the high density of OOV terms for this region is due to a confident use of local varieties by their speakers.

On the other hand, regions from Table 3 in which languages and dialects are spoken the least (i.e., EMI, LOM, TOS; cf. Figure 3, *bottom*) show a higher fraction of non-content tokens, a sign of the increasingly restricted function of language varieties. As prototypical examples, we can find "*cinno*" (EMI; en: *kid*), "*ratt*" (LOM; en: *rat(s)*), and "*diaccio*" (TOS; en: *icy/frozen/very cold*).

Exceptions on the ends are represented by PUG and LAZ (cf. Figure 3, *mid-top* and *mid-bottom*, respectively). PUG exhibits both content and non-content tokens, e.g., "*lu*" (en: *the* [M. SG.]), "*ientu*" (en: *wind*), and "*trmon*" (en: *stupid*), whereas LAZ only comprises non-content tokens, e.g., "*na*" (en: *a/an* [F. SG.]), "*de*" (en: *of*), and "*pe*" (en: *for*). While for PUG this can be ascribable to the small size of its subset and thus to the diversity of language included in it, the situation of LAZ is to be considered an outlier. Specifically, varieties spoken in LAZ are highly used indeed, but they

[13] This allows us to ground the discussion based on the subsets for which the PMI-based computation (Fano, 1961) behind the lexical artifacts package is more reliable.

are considered "ways of speaking" or "accents" of Standard Italian rather than proper language varieties (De Mauro, 1989). This has probably had an impact on the results of the aforementioned survey by ISTAT (2017) and justifies this divergence.

4 Experiments

In this section we present our experiments on the DIATOPIT corpus. Our objective is to understand how difficult it is to model diatopic language variation in Italy, i.e., by identifying coarse- and fine-grained geographical areas of a post based solely on its textual content.[14] Ultimately, this will help in building tools to reliably identify content for language varieties of Italy from social media, and thus to better represent them in NLP. We first introduce the experimental setup (Section 4.1) and the baselines we employed (Section 4.2). Then, we present the results and provide a discussion (Section 4.3).

4.1 Experimental Setup

Tasks We cast the problem of identifying the area from which a tweet has been posted into two tasks of increasing complexity: *(i) coarse-grained geolocation* (CG), i.e., predict the administrative region from which a tweet has been posted (classification task), and *(ii) fine-grained geolocation* (FG), i.e., predict latitude and longitude coordinates for the post (double-regression task). For each task we provide several experimental baselines (Section 4.2).

Data splits For training and testing the models, we divide the corpus into `train`, `dev`, and `test` sets. Given the highly-unbalanced distribution of instances across regions (cf. Table 2), for `dev` and `test` sets we draw a number of posts per region according to a smoothed distribution. Specifically, for each region r we take its raw number of instances I_r and we calculate a smoothed value $\sqrt{I_r}$, further adjusted by a multiplication factor λ to control the proportional size of the resulting `dev` and `test` sets.[15] This ensures a more reliable evaluation due to a higher percentage of instances in `dev` and `test` sets for under-represented regions. Moreover, we deliberately avoid sampling those instances at random, since this process could lead to a limited coverage of linguistic phenomena and microvariation in `dev` and `test`. We instead ask curators to manually select `dev` and `test` instances from a 50% random sample for each region[16] to be as representative as possible of a wide range of linguistic phenomena and microvariation. Additionally, we also ask them not to include instances that explicitly cite others (e.g., "*as my grandma says: 'X'*") to focus our evaluation on actual language use. Once the predefined smoothed value for each region was reached, we added the rest of the examples to the remaining 50% (i.e., `train`). Due to the very low number of instances for some regions, and thus scarcity of data for properly evaluating those, we decided to keep posts for the top-13 regions ($\geq$ 200 instances) for development and the top-17 regions ($\geq$ 50 instances) for testing (cf. Table 2), while leaving all 20 regions for training. This led to 13,669 examples for `train`, 552 examples for `dev`, and 818 examples for `test`, distributed as shown in Appendix D.

Evaluation metrics Since the distribution of instances per region is highly imbalanced, for the CG task we use macro-averaged scores so that each region in the evaluation set (either `dev` or `test`) is factored equally into the metric. Specifically, we employ macro-averaged precision (P), recall (R), and F_1 score. For the FG task, we instead use the mean error of the predicted coordinates from actual coordinates in kilometers (km), calculated using the Haversine formula.[17]

4.2 Baseline Models

We use several baseline models in order to provide reference points for future work using our corpus.

Naïve baselines For task CG we use a most-frequent baseline that always predicts the most frequent region in the training set (i.e., LAZ). For the FG task we instead employ a centroid baseline that computes the center point from training instances and predicts it for all test instances.

[14] This is in contrast to standard language/dialect identification tasks, in which the goal is to categorize texts into uniform language/dialect categories rather than identify areas where those are spoken – thus taking microvariation into account. Our formulation also differs from the Italy's language and dialect identification task (Aepli et al., 2022), in that we also deal with naturally occurring code-switched content and regional varieties of Standard Italian. Moreover, we model language from social media which is more spontaneous and does not necessarily adhere to orthography standards.

[15] We use $\lambda = 1.50$ and $\lambda = 2.25$ for `dev` and `test`, respectively, i.e., making the size of the `test` $3/2$ that of `dev`. For regions for which instances are extremely scarce, we simply draw the same number of `dev` instances for the `test` portion.

[16] This further ensures that `train` is not deprived of important signal since it was left untouched in this process.

[17] https://github.com/mapado/haversine

Machine learning models For both tasks we train two traditional models: for the CG task, we train a logistic regression (LR) and a support vector machine (SVM) classifier, whereas for FG we train a regression model based on *k*-nearest neighbors (*k*NN) and a decision tree (DT) regressor. We use the `scikit-learn`[18] count vectorizer for feature extraction and employ default hyperparameters.

Pretrained language models We fine-tune two monolingual and two multilingual transformer-based (Vaswani et al., 2017) models for each task. The monolingual models we use are AlBERTo (Polignano et al., 2019), a BERT-based (Devlin et al., 2019) model pre-trained on Italian text data from Twitter, and UmBERTo (Parisi et al., 2020), a RoBERTa-based (Liu et al., 2019) model pre-trained on the Italian portion of the OSCAR web-crawled corpus (Suárez et al., 2019). While DIATOPIT comprises *non*-Standard Italian content, we hypothesize that the pre-training material that has been used by those models (i.e., social media texts and raw data) may include content in language varieties of Italy due to the over-prediction of Italian of current language identifiers (cf. Section 2.1).

The multilingual models we use are instead multilingual BERT base (mBERT; Devlin et al., 2019), which is pre-trained on Wikipedia texts in 104 languages, and XLM-Roberta base (XLM-R; Conneau et al., 2020), which is pre-trained on the filtered CommonCrawl raw corpus in 100 languages. In addition to Italian, mBERT pre-training material includes Wikipedia content for some language varieties represented in DIATOPIT, i.e., Lombard [`lmo`], Piedmontese [`pms`] and Sicilian [`scn`], albeit it reflects an artificial use of language (Ramponi, 2022).

We use default Huggingface (Wolf et al., 2020) `TrainingArguments` hyperparameters, setting the learning rate to $5e^{-5}$ and training models for 10 epochs. For the CG task we use a batch size of 32 and cross-entropy loss, whereas for the FG task we train models using a batch size of 64 and mean squared error (MSE) loss. We use MSE loss instead of mean absolute error (MAE) loss as it assigns higher penalties to large errors.

4.3 Results and Discussion

In this section we report the results obtained by our baselines for both tasks. Results are averaged across 5 runs using different random seeds for shuffling the data and initializing the models.

[18] https://scikit-learn.org/stable/index.html

Method	P	R	F_1
Most frequent	$4.47_{\pm 0.0}$	$21.15_{\pm 0.0}$	$7.38_{\pm 0.0}$
LR	$60.36_{\pm 0.0}$	$45.92_{\pm 0.0}$	$49.29_{\pm 0.0}$
SVM	$63.83_{\pm 0.0}$	$51.04_{\pm 0.0}$	$53.95_{\pm 0.0}$
AlBERTo	$62.52_{\pm 2.3}$	$56.98_{\pm 1.2}$	$\mathbf{58.43}_{\pm 1.5}$
UmBERTo	$58.97_{\pm 2.4}$	$55.86_{\pm 2.2}$	$56.19_{\pm 2.2}$
mBERT	$59.71_{\pm 3.1}$	$56.48_{\pm 2.2}$	$57.29_{\pm 2.4}$
XLM-R	$57.73_{\pm 3.0}$	$51.35_{\pm 1.3}$	$51.86_{\pm 1.9}$

Table 4: Test set results for the CG task. We report average precision (P), recall (R), and macro F_1 scores across 5 runs ($\pm$: std dev). Best results are in bold.

4.3.1 Coarse-Grained Geolocation

Results on the CG task are presented in Table 4. The best-performing baseline is AlBERTo, with a macro F_1 score of 58.43, while – besides the most frequent baseline – the lowest score is obtained by LR, with a macro F_1 score of 49.29. Interestingly, the SVM classifier is a strong baseline even though it is far less computationally expensive than transformer-based models, performing better (+2.09) than XLM-R. A potential reason for traditional models to be competitive against large language models (LLMs) is that the variation of lexical items across varieties makes them very informative features. Furthermore, LLMs could suffer from suboptimal subword tokenization, given that tokenizers for these models are not optimized for the language varieties in our corpus. Overall, it appears that transformer-based models might benefit from being trained on in-domain data (i.e., Twitter for AlBERTo) or data containing a subset of the varieties represented in DIATOPIT (e.g., mBERT).

The CG task is generally challenging, not only because it represents a very unbalanced multi-class classification problem (cf. Table 2), but also because there are some language varieties that are very close across regions, especially in border areas. In Figure 4 we present the confusion matrix for our best-performing baseline (i.e., AlBERTo), showing the effect of these challenges on model predictions. For instance, the high class imbalance causes the model to perform better (especially with regards to recall) on highly represented regions (e.g., LAZ and CAM), while regions with a lower percentage of instances in the corpus tend to be predicted less frequently. Specifically, regions that are scarcely represented in training data are often confused with neighboring regions and/or regions where a simi-

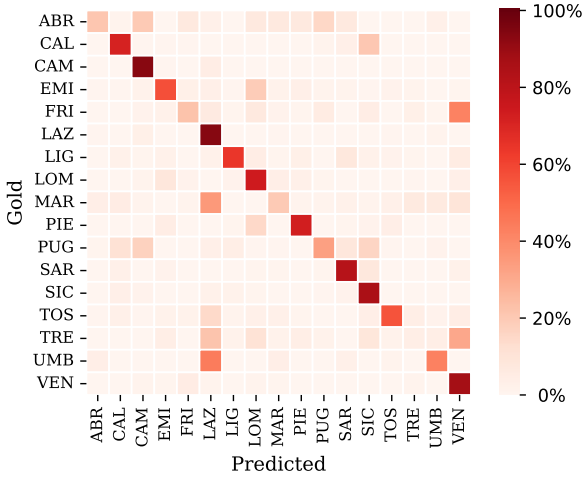


Figure 4: Confusion matrix for AlBERTo on the CG test set. Each row is normalized so that its sum is 100%.

Method	Avg dist (km)
Centroid	$281.04_{\pm 0.0}$
kNN	$245.60_{\pm 0.0}$
DT	$152.45_{\pm 1.4}$
AlBERTo	$\mathbf{151.54}_{\pm 7.8}$
UmBERTo	$207.65_{\pm 41.3}$
mBERT	$211.51_{\pm 39.4}$
XLM-R	$266.32_{\pm 23.8}$

Table 5: Test set results for the FG task. We report the average distance in kilometers across 5 runs (±: std dev). Best results are in bold (the lower, the better).

lar variety is spoken. This is the case of e.g., FRI and TRE, in which varieties of Venetian [vec] are spoken (amongst others), and thus instances are often misclassified as VEN, the region in which vec is predominantly used. Similarly, PUG is often confused with CAM, but also with SIC, despite not being near to it. This is because of language varieties spoken in the southern part of PUG (i.e., *Salentino* varieties), which are close to those of SIC, being both part of extreme southern varieties (cf. Pellegrini (1977) for more details). Results by region for all methods are in Appendix D.

Despite the aforementioned challenges, in part due to the simplification entailed in framing diatopic variation across space as a classification task in which the labels are administrative regions, the error analysis shows that models tend to confound regions that actually share common linguistic traits. This seems to indicate that DIATOPIT does reflect the actual distribution of language varieties in Italy.

4.3.2 Fine-Grained Geolocation

Results on the FG task for all baselines are presented in Table 5. Similarly to the coarse-grained geolocation task, the best-performing model is AlBERTo, with a mean average error of 151.54 km. Interestingly, DT performs similarly to AlBERTo (152.45 km; +0.91), even though it requires a fraction of the computational cost. Other transformer-based models have much higher error rates than AlBERTo, as well as a very large standard deviation across runs. This indicates that they are not sufficiently robust for modeling fine-grained geolocation. We hypothesize that the stability of results by AlBERTo compared to UmBERTo, mBERT, and XLM-R is due to the in-domain nature of textual data used during pre-training. Moreover, the good results obtained by DT suggest that current transformer-based models are rather limited for modeling language variation over space in highly multilingual areas such as Italy due to an insufficient vocabulary coverage. In future work we plan to experiment with token-free models (Xue et al., 2022; Clark et al., 2022; Tay et al., 2022) to assess if the vocabulary issue can be mitigated.

More generally, the improvement achieved by our best baseline over the centroid baseline for the FG task is comparable or better than the improvements obtained by the best-performing models in the Social Media Variety Geolocation (SMG) task at the 2020 VarDial Evaluation Campaign (Gaman et al., 2020), focused on the geolocation of social media posts in different geographical areas. While our best model's mean error improves by 46.08% over the centroid baseline, the models in the SMG task showed mean error improvements over the centroid baselines of 40.41%, 16,96%, and 47.97%.

5 Conclusion

We present DIATOPIT, the first corpus focused on diatopic variation in Italy for language varieties other than Standard Italian. Our analyses and experiments show that DIATOPIT is highly representative of actual use of Italy's language varieties, and can thus be used to advance research in the area. We plan to study divergences in orthography and code-switching in future work, in order to further assess vitality across varieties. Data and relevant materials (e.g., search terms) are available to the research community at `https://github.com/dhfbk/diatopit`.

Ethics Statement and Limitations

We release the corpus in the form of tweet IDs to be hydrated, in compliance to the Twitter developer policy. The corpus contains content that may be offensive or upsetting due to the occasional use of swear words by users. Latitude and longitude coordinates do not correspond to specific places within cities, but instead represent cities as a whole (i.e., posts within the same city have the same coordinates). Curators are part of the authors of this paper, and did the curation as part of their work. The corpus is meant to study diatopic language variation in Italy and can be used for research purposes only.

DIATOPIT includes content in regional varieties of Standard Italian as well as content written in the following local language varieties (ISO 639-3): `egl`, `fur`, `lij`, `lmo`, `nap`, `pms`, `rgn`, `scn`, `sdc`, `sdn`, `srd`, and `vec`, albeit with different amounts of data. Rare instances for `aae`, *Algherese Catalan* and *Calabrian Greek* are also present. Germanic varieties (e.g., `cim`, `mhn`, `wae`, *South Tyrolean*), `frp`, `lld`, and `svm` are instead mostly absent due to either the very low number of speakers or the sampling procedure. As regards to the latter, we plan to further extend the corpus with relevant samples classified as other than `it` by the Twitter language identifier to further mitigate under-representation of certain language varieties due to orthographic reasons or language branch.

References

Noëmi Aepli, Antonios Anastasopoulos, Adrian-Gabriel Chifu, William Domingues, Fahim Faisal, Mihaela Gaman, Radu Tudor Ionescu, and Yves Scherrer. 2022. Findings of the VarDial evaluation campaign 2022. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–13, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Francesco Avolio. 2009. *Lingue e Dialetti d'Italia*. Le Bussole. Carocci, Roma, Italy.

Matteo Bartoli, Ugo Pellis, and Lorenzo Massobrio. 1995. *Atlante Linguistico Italiano*. Istituto Poligrafico e Zecca dello Stato, Roma, Italy.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Massimo Cerruti and Riccardo Regis. 2005. 'Code switching' e teoria linguistica: La situazione italoromanza. *Italian Journal of Linguistics*, 17(1):179.

Bharathi Raja Chakravarthi, Gaman Mihaela, Radu Tudor Ionescu, Heidi Jauhiainen, Tommi Jauhiainen, Krister Lindén, Nikola Ljubešić, Niko Partanen, Ruba Priyadharshini, Christoph Purschke, Eswari Rajagopal, Yves Scherrer, and Marcos Zampieri. 2021. Findings of the VarDial evaluation campaign 2021. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–11, Kiyv, Ukraine. Association for Computational Linguistics.

Jonathan H. Clark, Dan Garrette, Iulia Turc, and John Wieting. 2022. Canine: Pre-training an efficient tokenization-free encoder for language representation. *Transactions of the Association for Computational Linguistics*, 10:73–91.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Tullio De Mauro. 1989. *Il romanesco ieri e oggi*. Bulzoni, Roma, Italy.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jacob Eisenstein. 2013. What to do about bad language on the internet. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 359–369, Atlanta, Georgia. Association for Computational Linguistics.

Robert M Fano. 1961. Transmission of information: A statistical theory of communications. *American Journal of Physics*, 29:793–794.

Mihaela Gaman, Dirk Hovy, Radu Tudor Ionescu, Heidi Jauhiainen, Tommi Jauhiainen, Krister Lindén, Nikola Ljubešić, Niko Partanen, Christoph Purschke, Yves Scherrer, and Marcos Zampieri. 2020. A report on the VarDial evaluation campaign 2020. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–14, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Bo Han, Afshin Rahimi, Leon Derczynski, and Timothy Baldwin. 2016. Twitter geolocation prediction shared task of the 2016 workshop on noisy user-generated text. In *Proceedings of the 2nd Workshop on Noisy*

User-generated Text (WNUT), pages 213–217, Osaka, Japan. The COLING 2016 Organizing Committee.

ISTAT. 2017. L'uso della lingua italiana, dei dialetti e di altre lingue in Italia. https://www.istat.it/it/archivio/207961. Accessed: 2023-02-01.

Karl Jaberg, Jakob Jud, and Glauco Sanga. 1987. *Atlante Linguistico ed Etnografico dell'Italia e della Svizzera Meridionale*, Italian edition. Unicopli, Milano, Italy.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Martin Maiden and Mair Parry. 1997. *The Dialects of Italy*. Routledge, London, England.

Christopher Moseley. 2010. *Atlas of the World's Languages in Danger*, 3rd edition. Memory of Peoples. UNESCO Publishing, Paris, France.

Loreto Parisi, Simone Francia, and Paolo Magnani. 2020. UmBERTo: an Italian language model trained with whole word masking. https://github.com/musixmatchresearch/umberto. Accessed: 2023-02-01.

Giovan Battista Pellegrini. 1977. *Carta dei Dialetti d'Italia*. Profilo dei Dialetti Italiani. Pacini, Pisa, Italy.

Marco Polignano, Pierpaolo Basile, Marco de Gemmis, Giovanni Semeraro, and Valerio Basile. 2019. AlBERTo: Italian BERT language understanding model for NLP challenging tasks based on tweets. In *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019)*, volume 2481. CEUR.

Alan Ramponi. 2022. NLP for language varieties of Italy: Challenges and the path forward. *arXiv preprint arXiv:2209.09757*.

Alan Ramponi and Sara Tonelli. 2022. Features or spurious artifacts? data-centric baselines for fair and robust hate speech detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3027–3040, Seattle, United States. Association for Computational Linguistics.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*. Leibniz-Institut für Deutsche Sprache.

Yi Tay, Vinh Q Tran, Sebastian Ruder, Jai Gupta, Hyung Won Chung, Dara Bahri, Zhen Qin, Simon Baumgartner, Cong Yu, and Donald Metzler. 2022. Charformer: Fast character transformers via gradient-based subword tokenization. In *The Tenth International Conference on Learning Representations, ICLR 2022*, Virtual.

Rob van der Goot, Alan Ramponi, Arkaitz Zubiaga, Barbara Plank, Benjamin Muller, Iñaki San Vicente Roncal, Nikola Ljubešić, Özlem Çetinoğlu, Rahmad Mahendra, Talha Çolakoğlu, Timothy Baldwin, Tommaso Caselli, and Wladimir Sidorenko. 2021. MultiLexNorm: A shared task on multilingual lexical normalization. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 493–509, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Appendix

A Data Statements

We present the data statements (Bender and Friedman, 2018) for DIATOPIT in the following.

CURATION RATIONALE. DIATOPIT consists of social media posts (partially and fully) written in language varieties of Italy other than Standard Italian, and is thus meant to encourage research on diatopic variation in Italy, study code-switching and divergences in orthography for local language varieties, and serve as a basis for responsible development of annotated resources for Italy's varieties. Details on corpus creation are given in Section 2.

LANGUAGE VARIETIES. The corpus includes content in regional varieties of Standard Italian (`ita`), as well as content written in the following local language varieties (ISO 639-3 codes, wherever available): `egl`, `fur`, `lij`, `lmo`, `nap`, `pms`, `rgn`, `scn`, `sdc`, `sdn`, `srd`, and `vec`, albeit with different amounts of data. Rare instances for `aae`, *Algherese Catalan* and *Calabrian Greek* are also present. Orthographic variation is common due to the spontaneous written speech of social media posts and the lack of standardization of most language varieties.

SPEAKER DEMOGRAPHIC. The corpus consists of anonymized social media posts, and thus user demographics are not known.

ANNOTATOR DEMOGRAPHIC. Two curators native to Italy with good knowledge of Italy's language varieties and background in NLP and sociolinguistics. They identify themselves as a woman and a man, with age ranges 20–30 and 30–40, and native speakers of `ita`, `srd`, and `vec`. Additional native speakers who have been consulted during curation in the presence of doubtful cases greatly vary in terms of demographic characteristics.

SPEECH SITUATION AND TEXT CHARACTERISTICS. The interaction is mainly asynchronous and the intended audience is everyone. The modality is (spontaneous) written text, the genre is social media without any particular topical focus due to the sampling procedure (cf. Section 2). Social media posts have been produced between `2020-07-01` and `2022-06-30`, and collected in September 2022.

PREPROCESSING AND DATA FORMATTING. All posts have been anonymized by replacing user mentions, email addresses and URLs with placeholders (i.e., `[USER]`, `[EMAIL]` and `[URL]`, respectively). Additionally, explicit location mentions derived from cross-posting have been replaced with the `[LOCATION]` placeholder. Newline characters have been replaced with single spaces. Latitude and longitude coordinates have been computed by taking the central point from the 4-point bounding box of city areas as provided by the Twitter APIs.

B Corpus Augmentation

Step 1 Data augmentation for geographical regions with $\leq 1\%$ instances I over the total has been carried out based on their initial amount of data (cf. Table 6, *top*). For regions with $I < 0.5\%$ posts (i.e., severely under-represented), all the posts matching at least an OOV token have been manually curated for inclusion (N = 4,606). For regions with $0.5\% \leq I \leq 1.0\%$ posts (i.e., moderately under-represented), a random 10% of the posts matching at least an OOV token have been manually curated for inclusion (N = 6,107). This led to 718 extra posts across all those regions, and notably an increment of more than $2\times$ instances for some regions (e.g., EMI: $0.99\% \rightarrow 2.41\%$; FRI: $0.70\% \rightarrow 1.70\%$; LIG: $0.62\% \rightarrow 1.37\%$).

Step 2 All regions except the over-represented LAZ and CAM (i.e., those with $I \leq 20.0\%$ posts over the total) were used to calculate highly-discriminative tokens for further sampling of posts (cf. Table 6, *bottom*). This led to N = 4,384 social media posts, 1,961 of which have been included in the final corpus after curation.

C Details about the Correlation Analysis

For the correlation analysis in Section 3.3 we took data from Table 1 of the survey by ISTAT (2017) on the usage of languages and dialects across Italy's administrative regions. Specifically, for our calculation we relied on percentages indicating the use of languages and dialects with friends, which is typically the case for spontaneous and informal social media content that includes local language varieties of Italy. Nevertheless, we found a similar correlation when considering the family context.

D Additional Details on the Experiments

The distribution of instances for the experiments is in Table 7, whereas results for the CG task divided by region and method are presented in Table 8.

Step	I (%)	Regions (relative percentage)
1	$[0.5\%, 1.0\%]$	EMI (0.99%), MAR (0.90%), ABR (0.85%), PIE (0.82%), FRI (0.70%), LIG (0.62%)
	$< 0.5\%$	TRE (0.23%), BAS (0.19%), MOL (0.15%), VAL (0.03%)
2	$\leq 20.0\%$	VEN (4.19%), LOM (3.79%), SIC (3.08%), TOS (2.56%), EMI (2.41%), PUG (1.86%), FRI (1.70%), CAL (1.57%), SAR (1.51%), PIE (1.49%), LIG (1.37%), MAR (1.35%), ABR (1.11%), UMB (1.09%), TRE (0.39%), BAS (0.35%), MOL (0.25%), VAL (0.09%)

Table 6: Geographical regions (and their relative percentages at the beginning of each stage) that have been selected for the two steps of data augmentation, i.e., step 1 (*top*) and step 2 (*bottom*).

ABR	BAS	CAL	CAM	EMI	FRI	LAZ
151 / – / 15	49 / – / –	282 / 27 / 27	3,027 / 85 / 128	320 / 30 / 45	220 / 25 / 25	5,607 / 115 / 173
LIG	**LOM**	**MAR**	**MOL**	**PIE**	**PUG**	**SAR**
223 / 25 / 25	696 / 43 / 64	181 / – / 16	35 / – / –	238 / 25 / 25	266 / 27 / 27	362 / 31 / 47
SIC	**TOS**	**TRE**	**UMB**	**VAL**	**VEN**	
620 / 40 / 60	421 / 34 / 51	52 / – / 9	136 / – / 14	14 / – / –	769 / 45 / 67	

Table 7: Distribution of `train` / `dev` / `test` instances by region for the sake of computational experiments.

Region		Method					
Abbr.	Full name	LR	SVM	AlBERTo	UmBERTo	mBERT	XLM-R
ABR	*Abruzzo*	$0.00_{\pm 0.0}$	$21.05_{\pm 0.0}$	$27.28_{\pm 14.7}$	$31.06_{\pm 12.4}$	$44.28_{\pm 10.1}$	$15.95_{\pm 4.7}$
CAL	*Calabria*	$61.90_{\pm 0.0}$	$57.14_{\pm 0.0}$	$67.22_{\pm 2.2}$	$56.98_{\pm 8.5}$	$58.08_{\pm 5.0}$	$41.42_{\pm 6.7}$
CAM	*Campania*	$80.14_{\pm 0.0}$	$81.75_{\pm 0.0}$	$89.52_{\pm 1.7}$	$91.02_{\pm 1.1}$	$89.68_{\pm 1.5}$	$89.73_{\pm 1.4}$
EMI	*Emilia Romagna*	$47.06_{\pm 0.0}$	$55.26_{\pm 0.0}$	$62.04_{\pm 6.4}$	$63.18_{\pm 2.6}$	$56.60_{\pm 4.9}$	$56.88_{\pm 2.7}$
FRI	*Friuli-Venezia Giulia*	$36.36_{\pm 0.0}$	$30.00_{\pm 0.0}$	$28.62_{\pm 4.5}$	$36.81_{\pm 5.0}$	$25.24_{\pm 4.9}$	$24.78_{\pm 8.5}$
LAZ	*Lazio*	$72.29_{\pm 0.0}$	$78.47_{\pm 0.0}$	$87.47_{\pm 0.5}$	$88.95_{\pm 1.4}$	$85.87_{\pm 0.8}$	$87.01_{\pm 1.3}$
LIG	*Liguria*	$48.65_{\pm 0.0}$	$66.67_{\pm 0.0}$	$68.95_{\pm 4.8}$	$69.72_{\pm 5.2}$	$76.84_{\pm 2.6}$	$78.22_{\pm 1.8}$
LOM	*Lombardia*	$59.84_{\pm 0.0}$	$60.80_{\pm 0.0}$	$70.06_{\pm 1.7}$	$72.44_{\pm 4.8}$	$71.97_{\pm 1.9}$	$70.70_{\pm 3.2}$
MAR	*Marche*	$26.09_{\pm 0.0}$	$25.00_{\pm 0.0}$	$20.96_{\pm 5.0}$	$21.57_{\pm 10.5}$	$25.75_{\pm 5.7}$	$14.21_{\pm 5.8}$
PIE	*Piemonte*	$75.56_{\pm 0.0}$	$74.51_{\pm 0.0}$	$73.21_{\pm 3.7}$	$65.46_{\pm 5.6}$	$71.70_{\pm 1.5}$	$65.48_{\pm 6.8}$
PUG	*Puglia*	$40.00_{\pm 0.0}$	$38.89_{\pm 0.0}$	$41.33_{\pm 5.3}$	$39.97_{\pm 6.0}$	$37.13_{\pm 7.6}$	$29.07_{\pm 5.3}$
SAR	*Sardegna*	$78.16_{\pm 0.0}$	$80.95_{\pm 0.0}$	$80.91_{\pm 3.8}$	$80.19_{\pm 2.5}$	$80.45_{\pm 3.1}$	$76.53_{\pm 2.6}$
SIC	*Sicilia*	$74.38_{\pm 0.0}$	$74.80_{\pm 0.0}$	$78.42_{\pm 2.3}$	$79.76_{\pm 2.6}$	$82.13_{\pm 3.8}$	$78.82_{\pm 3.2}$
TOS	*Toscana*	$62.50_{\pm 0.0}$	$74.23_{\pm 0.0}$	$67.36_{\pm 3.4}$	$70.71_{\pm 1.1}$	$69.28_{\pm 3.4}$	$68.37_{\pm 4.5}$
TRE	*Trentino-Alto Adige*	$0.00_{\pm 0.0}$	$0.00_{\pm 0.0}$	$4.72_{\pm 6.5}$	$0.00_{\pm 0.0}$	$10.30_{\pm 15.1}$	$0.00_{\pm 0.0}$
UMB	*Umbria*	$0.00_{\pm 0.0}$	$23.53_{\pm 0.0}$	$46.75_{\pm 7.2}$	$5.17_{\pm 7.1}$	$10.20_{\pm 11.8}$	$7.11_{\pm 10.2}$
VEN	*Veneto*	$75.00_{\pm 0.0}$	$74.17_{\pm 0.0}$	$78.38_{\pm 2.4}$	$82.32_{\pm 2.3}$	$78.35_{\pm 3.6}$	$77.29_{\pm 1.5}$

Table 8: Test set results for the CG task by region. We report average macro F_1 scores across 5 runs ($\pm$: std dev).