Language Varieties of Italy: Technology Challenges and Opportunities

Alan Ramponi
Fondazione Bruno Kessler
Trento, Italy
alramponi@fbk.eu

Abstract

Italy is characterized by a one-of-a-kind linguistic diversity landscape in Europe, which implicitly encodes local knowledge, cultural traditions, artistic expressions, and history of its speakers. However, most local languages and dialects in Italy are at risk of disappearing within a few generations. The NLP community has recently begun to engage with endangered languages, including those of Italy. Yet, most efforts assume that these varieties are *under-resourced* language monoliths with an established written form and homogeneous functions and needs, and thus highly interchangeable with each other and with *high-resource*, standardized languages. In this paper, we introduce the linguistic context of Italy and challenge the default *machine-centric* assumptions of NLP for Italy's language varieties. We advocate for a shift in the paradigm from *machine-centric* to *speaker-centric* NLP, and provide recommendations and opportunities for work that prioritizes languages and their speakers over technological advances. To facilitate the process, we finally propose building a local community towards responsible, participatory efforts aimed at supporting vitality of languages and dialects of Italy.

1 Introduction

> ''*Italy holds especial treasures for linguists. There is probably no other area in Europe in which such a profusion of linguistic variation is concentrated into so small a geographical area.*''
>
> — Maiden and Parry (1997)

Language is a primary means for communication that is intrinsic to the expression of culture. Through languages, we signal our social identities and convey part of our heritage (Thomason, 2015). However, according to the UNESCO *Atlas of World's Languages in Danger* (Moseley, 2010) about half of the spoken languages in the world are at risk of disappearing by the end of the century. Ultimately, this will lead to a loss of an integral part of cultures and traditions (Hale et al., 1992).

The natural language processing (NLP) community has recently started to include endangered languages in its repertoire, and language varieties of Italy are no exception. However, most of the efforts in NLP implicitly assume that these language varieties are just *under-resourced* entities (in terms of *written* data availability) with an established written form, and with the same functions and technological needs of *high-resource* standardized languages with institutional support, such as Italian or English (Bird, 2022). This *machine-centric* approach not only fails to acknowledge that most endangered languages are primarily oral, without a standardized orthography and canonical variant, often code-switched with a co-territorial ''high-prestige'' standardized language, and serving different language functions to other languages within the local linguistic ecosystem (Fishman, 2001), but also disregards *what* and *how* technologies should be built to safeguard endangered languages in the interest of speech communities (Bird, 2020; Caselli et al., 2021).

In this paper, we discuss the technology challenges and opportunities for **language varieties of Italy**, one of the most linguistically diverse landscapes in Europe which, according to UNESCO (Moseley, 2010), currently counts over 30 languages in danger. Italy's languages and dialects are not only many for such a small area (Maiden and Parry, 1997), but they are also very different from each other, and their linguistic distance does not typically relate to geographical distance (Avolio, 2009). Most of these varieties are Romance, albeit Germanic, Slavic, Albanian, and Hellenic ones also shape the Italian linguistic landscape. As for the majority of endangered languages,

Transactions of the Association for Computational Linguistics, vol. 12, pp. 19–38, 2024. https://doi.org/10.1162/tacl_a_00631
Action Editor: Saif Mohammad. Submission batch: 6/2023; Revision batch: 9/2023; Published 1/2024.
© 2024 Association for Computational Linguistics. Distributed under a CC-BY 4.0 license.

most of Italy's language varieties comprise many local variants, have no standardized written form, and are just occasionally written, insofar as they are primarily used in spoken, informal settings. They typically exist in a peculiar diglossic situation with Italian, and vary in terms of recognition, protection, economic incentives, and prospects.

After introducing the linguistic situation in Italy (Section 2), we review efforts in NLP for its languages and dialects (Section 3). We then discuss the *machine-centric* assumptions of the default NLP approach when dealing with these varieties, namely, the exaggerated focus on ''machine-readable'' written data, the little regard for the representativeness of such materials of speech communities, and the homogeneous view of functions, uses, and needs across language varieties (Section 4). We argue that language varieties of Italy should not be approached as a *data commodity* for machine learning advances, and that technology should serve language varieties and their speakers and not the other way round. We thus present recommendations and opportunities for *speaker-centric* NLP and advocate for a local community aimed at responsibly supporting vitality of Italy's varieties through sensitization on ethical engagement, sharing of practices, participatory collaboration, and active awareness-raising (Section 5). Finally, we provide our conclusions (Section 6).

Contributions We *i)* expose the NLP community to endangered language varieties of Italy, *ii)* survey computational work for these varieties, and *iii)* shed light on the main assumptions and shortcomings of the standard machine-centric NLP approach. *iv)* We then identify directions and opportunities for responsible, speaker-centric efforts aimed at preserving language varieties of Italy. Finally, *v)* we call for a local, multidisciplinary community that supports participatory work and knowledge sharing towards common goals. We hope our recommendations will be useful for the safeguarding of other endangered languages, too.

2 Linguistic Context of Italy

2.1 History and Standard Italian

Italy is one of the most diverse landscapes in Europe in terms of language varieties (Avolio, 2009). Unified late, the country was previously a collection of states with their own local languages. After the political unification in 1861, Standard Italian (ISO 639-3 code: `ita`) was adopted by the state as the official language, making it a unifying element. Italian emerged from a literary language based on Vulgar Latin, and specifically from the Tuscan variety as spoken by the Florentine upper-class society (Maiden and Parry, 1997). At unification time, Italian was spoken by less than 10% of the population (De Mauro, 1963), and rates of literacy remained low for over a century, especially in rural areas. Along with education, the rise of mass media played a crucial role in establishing the widespread use of Standard Italian, mirrored by a substantial decline in the use of local languages.[1] Nowadays, Italian is the fourth most widely spoken Romance language in the world with about 68M speakers (Eberhard et al., 2022).

2.2 Languages and Dialects of Italy

Despite the establishment of Italian as national language, many local languages and dialects are still currently spoken in Italy. In Table 1 we report the language varieties of Italy classified as endangered by UNESCO (Moseley, 2010) along with their ISO 639-3 code (wherever available), linguistic branch, level of endangerment, number of speakers, and whether they have a standardized written form.

While most varieties have fewer than 1M speakers and are definitely or severely endangered, some are still used even by younger generations in informal settings, i.e., language varieties spoken in the south and northeast areas of the Italian peninsula (ISTAT, 2017). Just like most languages of the world, languages and dialects of Italy are primarily used in spoken contexts, and only a fraction of them have a recently established written form. Most language varieties of Italy are Romance, insofar as they locally evolved from Vulgar Latin like Standard Italian.[2] The rest include non-Latin linguistic minorities from Germanic, Albanian, Hellenic, and Slavic Indo-European branches.

Due to the complex historical and sociopolitical motivations behind the use—with negative connotation—of the term *dialetti* (''dialects'') for

[1] Estimates indicate that 45.9% of the population mainly speak Italian at home, 32.3% use Italian and a local language, and 14.1% mostly speak a local language (ISTAT, 2017).

[2] Indeed, in this context the frequently used ''*Italian* languages/dialects'' expression is a misnomer (Avolio, 2009).

Id	Name	Branch	LoE	Speakers	Id	Name	Branch	LoE	Speakers
nap	Neapolitan	Romance	○	6.6M	*roa-via*	Vivaro-Alpine Occitan	Romance	◉	65K
scn	Sicilian	Romance	○	4.7M	*roa-gis*	Gallo-Italic of Sicily	Romance	◉	60K
vec	Venetian◇	Romance	○	3.9M	lld	Ladin◇	Romance	◉	41K
lmo	Lombard	Romance	◉	3.5M	*grk-gri*	Griko	Hellenic	●	35K
egl	Emilian	Romance	◉	2.0M	*roa-alc*	Algherese Catalan	Romance	◉	34K
pms	Piedmontese◇	Romance	◉	1.4M	wae	Walser	Germanic	●	13K
rgn	Romagnol	Romance	◉	1.1M	mhn	Mòcheno	Germanic	◉	2K
srd	Sardinian◇*	Romance	◉	1.0M	*grk-cal*	Calabrian Greek	Hellenic	●	1K
fur	Friulian◇	Romance	◉	0.6M	*roa-fae*	Faetar	Romance	◉	1K
lij	Ligurian◇	Romance	◉	0.5M	svm	Molise Slavic	Slavic	●	1K
gem-sty	South Tyrolean	Germanic	○	0.3M	*sla-res*	Resian	Slavic	◉	<1K
aae	Arbëreshë Albanian	Albanian	◉	0.1M	cim	Cimbrian	Germanic	◉	<1K
sdn	Gallurese	Romance	◉	0.1M	*roa-gar*	Gardiol	Romance	●	<1K
sdc	Sassarese	Romance	◉	0.1M	itk	Judeo-Italian	Romance	◉	<1K
frp	Francoprovençal	Romance	◉	71K	*gem-toi*	Töitschu	Germanic	●	<1K

Table 1: Endangered language varieties of Italy. Levels of endangerment (LoE) are: ○ vulnerable, ◉ definitely endangered, and ● severely endangered (Moseley, 2010). Language identifiers (Id) follow ISO 639-3 codes, wherever available; if not, we use an arbitrary designator (*italicized*). The number of speakers is at a country level and is mainly taken from Glottolog and Ethnologue estimates. ◇The language variety has a standardized written form. *Sardinian is a macro-language that includes Logudorese (src) and Campidanese (sro). Notes: Romani (rom) and Corsican (cos) are not included due to low specificity to Italy, and for Bavarian (bar) and Alemannic (gsw) we keep the local variants that are spoken in Italy, i.e., South Tyrolean (gem-sty) and Walser (wae), respectively.

language varieties of Italy (Avolio, 2009), and the range of meanings that the term assumes according to the context in which it is situated (Berruto, 2005), we hereafter refer to those languages and dialects as *language varieties*.[3] In the following, we contextualize endangered language varieties of Italy (Table 1) within the linguistic macro-areas proposed in the renowned *Carta dei dialetti d'Italia* (Pellegrini, 1977). An indicative linguistic map is also shown in Figure 1. For more details on the features of each variety and a systematic characterization of them, including local variants, we refer the reader to relevant linguistic studies and overviews on the topic (Pellegrini, 1977; Maiden and Parry, 1997; Avolio, 2009, *inter alia*).

Cisalpine System This includes Gallo-Italic varieties situated in northern Italy (i.e., *Piedmontese*, *Ligurian*, *Lombard*, *Emilian*, *Romagnol*) and *Venetian*, along with their many local variants.

Friulian System *Friulian*, a Rhaeto-Romance language recognized by the Italian state and spoken in northeast Italy, along with its local variants.

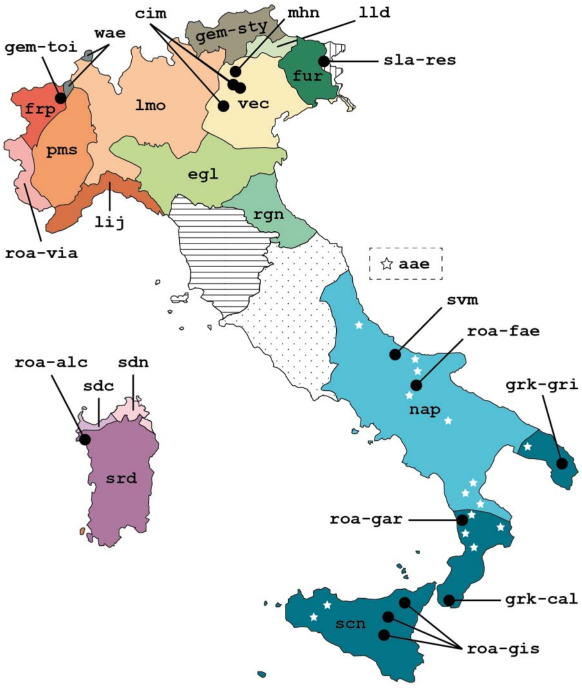

Figure 1: Map of Italy's endangered language varieties. Boundaries serve as a reading guide only: Italy's varieties often lie on a continuum without abrupt borders.

Tuscan System Non-endangered language varieties that are closely related to Standard Italian (middle-northern Italy; Figure 1, *horizontal lines*).

[3] The term prevents any judgment on the prestige status of each variety, and avoids discussions on political matters that are not the focus of this paper.

Middle-southern System Non-endangered varieties in central Italy (Figure 1, *dots*), intermediate-southern varieties (e.g., *Neapolitan* as a group of closely related varieties spoken in southern continental Italy), and extreme-southern varieties (i.e., *Sicilian*, its local variants, and related varieties).

Sardinian System Varieties spoken in the island of Sardinia. These include the officially recognized *Sardinian* macro-language (comprising *Logudorese* and *Campidanese*) and *Gallurese* and *Sassarese*, spoken in the north of the island.

Other Varieties These include protected varieties such as *Francoprovençal* as spoken in the Aosta Valley and Piedmont and the *Vivaro-Alpine Occitan* variety (all in northwest Italy), the Rhaeto-Romance *Ladin* language (northeast Italy), the Austro-Bavarian *South Tyrolean* variety (northern Italy), and Slovenian varieties (northeast Italy; Figure 1, *vertical lines*), including *Resian*. Varieties of *Judeo-Italian* are also spoken across the country by very small Jewish communities.

Language Enclaves A number of language islands enrich the already complex linguistic landscape of Italy (Figure 1, *black dots*). These include Germanic varieties in northern Italy (i.e., *Cimbrian*, *Mòcheno*, *Walser*, *Töitschu*); Modern Greek varieties in the Salento and Calabria areas, southern Italy (i.e., *Griko* and *Calabrian Greek*); the *Molise Slavic* Serbo-Croatian variety in the Molise region, middle-southern Italy; the Francoprovençal *Faetar* variety spoken in two small towns in Apulia and the Vivaro-Alpine *Gardiol* enclave in the Calabria region, southern Italy; the *Gallo-Italic of Sicily* Lombard enclave in the island of Sicily; the *Algherese Catalan* variant spoken in Alghero (Sardinia); and *Arbëreshë Albanian*, whose communities are scattered across southern Italy (Figure 1, *white stars*).

2.3 Regional Italian

Alongside Italian and indigenous language varieties and linguistic minorities, regional varieties of Standard Italian (hereafter, *regional Italian*) are also spoken by most Italian speakers. Varieties of regional Italian result from a geographical differentiation of Standard Italian after its widespread adoption, and differ from each other at various levels, i.e., syntax, morphology, phonetics, phonology, and prosody (Cerruti, 2011; Avolio, 2009). The various forms of regional Italian mostly match macro-linguistic areas of language varieties (cf. Section 2.2), and vary according to social and educational factors (Avolio, 2009).

3 Language Varieties of Italy and NLP

The study, preservation, and promotion of language diversity have recently gained increasing attention in the NLP community. Initiatives such as the ACL 2022 special theme on ``*Language Diversity: From Low-resource to Endangered Languages*'' (Muresan et al., 2022), the ACL special interest group SIGEL,[4] and relevant workshops, e.g., ComputEL (Harrigan et al., 2023), Eurali (Ojha et al., 2022), AmericasNLP (Mager et al., 2021), have been proposed. Moreover, the VarDial series of workshops (Scherrer et al., 2023, *inter alia*) is being routinely organized to promote the study of diatopic variation of language varieties and dialects.

In the following, we review previous work in NLP for Italy's varieties, from monolingual (Section 3.1) to multilingual efforts (Section 3.2), and highlight commonalities and differences in the shortcomings of both research lines (Section 3.3).

3.1 NLP for Specific Varieties of Italy

Natural language processing research for specific languages and dialects of Italy is scarce and scattered across disciplines. The most studied language variety is Venetian, for which there exist work on morphological analysis (Tonelli et al., 2010), part-of-speech tagging (POS; Jaber et al., 2011), word sense disambiguation (Conforti and Fraser, 2017), and a preliminary investigation on Venetian-English machine translation (MT; Delmonte et al., 2009). Ligurian has also recently gained attention in NLP, with work on text normalization (Lusito et al., 2023) and the development of a Universal Dependency (UD; de Marneffe et al., 2021) treebank for the *Genoese* variety (Lusito and Maillard, 2021). A small set of *Vivaro-alpine* examples has been included in an Occitan subcorpus with POS annotations (Bernhard et al., 2018, 2021), whereas for Ladin, previous work includes MT from and to Italian for the *Val Badia* variety (Frontull, 2022). MT has also been studied for Sicilian⇆English and zero-shot Sicilian⇆Italian (Wdowiak, 2022), and

[4] https://acl-sigel.github.io/.

for Italian→Sardinian (Tyers et al., 2017) and Catalan→Sardinian (Fronteddu et al., 2017).

Among severely endangered varieties, Griko is the most represented in NLP. Previous work includes two Griko-Italian parallel corpora: A corpus of narratives with POS annotations (Anastasopoulos et al., 2018; Chaudhary et al., 2021) and a small speech-derived corpus annotated with morphosyntactic, POS, glosses, and speech-related information (Boito et al., 2018; Lekakou et al., 2013). Other efforts in this space include Molise Slavic, for which field recordings, transcriptions, and Italian and German translations have been made available for the varieties of *Acquaviva Collecroce*, *San Felice*, and *Montemitro* (Breu, 2017).

A number of resources have been produced for plurilingualism areas of Italy where South Tyrolean is spoken, such as a multilingual corpus of computer-mediated communication (Frey et al., 2016), and a longitudinal trilingual corpus of young learners (Glaznieks et al., 2022). Preliminary efforts such as a morphosyntactic specification for Resian (Erjavec, 2017), a lexical database for Sardinian, Gallurese and Sassarese (Angioni et al., 2018), and a tagset for Cimbrian varieties (Agosti et al., 2012) have also been carried out. There also exist a few cultural institutes that have developed tools and resources that can be interrogated online, e.g., Micurá de Rü[5] (Ladin), and Kulturinstitut Lusérn[6] (Cimbrian), *inter alia*.

3.2 Varieties of Italy in Multilingual NLP

Language varieties of Italy are increasingly represented in multilingual research. Friulian, Ladin, Neapolitan, and Venetian have been included in the SIGMORPHON shared tasks on morphological inflection in 2018–2020 (Cotterell et al., 2018; McCarthy et al., 2019; Vylomova et al., 2020), though the latter two have been discontinuously represented. More recently, a language and dialect identification shared task has been proposed (Aepli et al., 2022), for which participants were given Wikipedia dumps of 11 varieties of Italy and were asked to classify text samples for a subset of the given varieties. Friulian, Ligurian, Lombard, Sicilian, Sardinian, and Venetian have also been included in a translation model covering 202 languages (NLLB Team et al., 2022), and a corpus for cross-lingual spoken language understanding has been annotated with slot and intent information in South Tyrolean and Neapolitan (van der Goot et al., 2021b; Aepli et al., 2023).

Other efforts including language varieties of Italy are sparse and mainly focus on learning methods, e.g., for learning contextualized cross-lingual word embeddings in low-resource scenarios (Griko in Wada et al., 2021), for language identification of text sequences in mixed-language documents (Lombard-English in King and Abney, 2013), or for investigating the effect of pretraining language selection on downstream zero-shot transfer (Piedmontese in Malkin et al., 2022).

Multilingual pretrained language models have also been proposed in recent times to widen language coverage in NLP, e.g., mBERT (Devlin et al., 2019), mBART (Liu et al., 2020), and XLM-R (Conneau et al., 2020). mBERT includes some of Italy's varieties, namely, Lombard, Piedmontese and Sicilian, albeit under-represented in terms of pretraining data compared to other languages. Training material is taken from entire Wikipedia editions, regardless of the covered variants, quality issues, and the representativeness of such language of speech communities (cf. Section 4.2).

3.3 NLP Serving Varieties of Italy or the Other Way Round?

From a closer look, we can observe that the attention to local language varieties and the very objectives of research efforts generally diverge between monolingual and multilingual NLP studies. Most efforts for specific language varieties of Italy are explicitly intended to study or support local languages and dialects, state the orthographies and local variants being considered, and are often conducted by members of the target speech communities—and are thus potentially driven by actual or perceived needs. On the other hand, recent trends in multilingual NLP are typically centered on computational advances (e.g., *scaling*, *generalizing*) rather than on varieties and their speakers. Indeed, most work in this space is driven by language technology agendas of standardized languages (Bird, 2022), and view the *under-resourcedness* of written content as a pivotal problem to be directly or indirectly fixed, do not mention which variants and orthographies of the language varieties have been included—and why they have been chosen over the others—

[5] `https://www.micura.it/`.

[6] `https://www.istitutocimbro.it/`.

perpetuating language monolithicity assumptions, and implicitly presume that language varieties are all the same in terms of functions, uses, and their speakers' needs (Section 4).

What both research strands have in common is that the active involvement of speech communities at various stages of the design process (e.g., to express needs or assess the envisioned technology rather than merely acting as *data producers*) is typically left unspecified. This confirms similar findings by Caselli et al. (2021) and motivates us to propose new ways of working centered on language varieties and their communities (Section 5).

4 The Default *Machine-centric* Approach

In this section, we provide an in-depth overview of the main assumptions and shortcomings of the default NLP approach—what we refer to as *machine-centric* NLP—with a focus on language varieties of Italy. We first discuss the persistent focus on written data scarcity and how this is often perceived as a problem to be solved (Section 4.1). Second, we focus on widespread text collections that are typically used for training language models, arguing that the common practice of *language data as a commodity* fails to represent language varieties and their speakers (Section 4.2). Lastly, we discuss the intrinsic assumptions of the standard approach, namely the lack of regard for functions, contexts and needs of varieties (Section 4.3).

4.1 Persistent Emphasis on Data Scarcity

A common argument in NLP work involving local language varieties of Italy is that these languages and dialects are *under-resourced* in terms of ‘‘machine-readable’’ written data, and are therefore in need of more resources—or computational means to bridge the gap—in order to take full advantage of language technologies. This view not only fails to acknowledge the reasons behind written data scarcity, but also implicitly homogenizes the contexts in which such language varieties are situated and the diverse aspirations for written text—and thus the needs for consequent technologies.

The focus on ‘‘data quantity’’ is widely rooted in the NLP community, and the amount of machine-readable language resources has also recently been used as a criterion for classifying the world's languages and highlighting their technological disparity. For instance, in the taxonomy of world's languages according to data availability by Joshi et al. (2020), 10 endangered varieties of Italy are in the second-worst position (1: *The Scraping-bys*), while the rest belong to the worst position (0: *The Left-behinds*).[7] Despite the best of intentions, this classification decouples the unique situation of each variety from its volume of machine-readable resources. While the amount of written data resources and the position in the ‘‘technologization race’’ are probably of interest to standardized and ‘‘*would-be* standardized languages’’[8] (see Bird, 2022), these factors are of little significance for most varieties, for which interests typically relate to culture preservation, language learning, and intergenerational transmission (Section 5).

Given the varied linguistic and socio-political contexts of Italy's varieties (Section 4.3), it is therefore more appropriate to outline *which* these resources are rather than *how many* they are. Hence, we extended the search by Joshi et al. (2020), which originally included the LDC catalog,[9] ELRA Map,[10] and Wikipedia, by covering additional repositories and all Italy's varieties. We searched for main and alternate names of each variety (e.g., `nap`: *Neapolitan*, *Neapolitan-Calabrese*, *Continental Southern Italian*) on OLAC (Simons and Bird, 2003), the CLARIN Virtual Language Observatory (Hinrichs and Krauwer, 2014), and OPUS (Tiedemann, 2012). The latter also includes data from educational resources, e.g., Tatoeba,[11] QED (Abdelali et al., 2014) and localization data from open-source software projects e.g., Ubuntu, Gnome, Mozilla. To further include language resources that have not been submitted to mainstream repositories, we also queried Google Scholar for publications that mention both NLP and a main or alternate name of a variety. We thoroughly inspected the top 50 results for each query and retained all entries that present or use language resources. Finally, we categorized all publicly available, curated language resources according to their language

[7] Either because explicitly indicated or not included at all.

[8] In the context of Italy, *would-be* standardized languages mostly match varieties in territories where bilingualism is officially granted by national or regional laws (Section 4.3).

[9] `https://catalog.ldc.upenn.edu/`.

[10] `https://catalog.elra.info/`.

[11] `https://tatoeba.org/`.

Corpus	Id	Genre	Annotation	Parallel	Size	Data
xSID (van der Goot et al., 2021b)	*gem-sty*	🔊	Slot; Intent	*multi*	800	URL
UoI (Boito et al., 2018)	*grk-gri*	🔊	Morph; POS; Glo; Sp	ita	330	URL
GrikoNarrative* (Anastasopoulos et al., 2018)	*grk-gri*	📕	POS	ita	942	URL
NormLigurian* (Lusito et al., 2023)	lij	📕 ♫ 🪄	Norm	*std*	4,394	URL
UD_Ligurian-GLT (Lusito and Maillard, 2021)	lij	📕 🪄 🔊 📰 W ☁	Morph; POS; Dep	—	316	URL
sid4lr (Aepli et al., 2023)	nap	🔊	Slot; Intent	*multi*	800	URL
Restaure (Bernhard et al., 2018, 2021)	*roa-via*	📕	POS	—	39	URL
Na-našu (Acquaviva) (Breu, 2017)	svm	🔊	Glo; Sp	deu;eng;ita	890	URL
Na-našu (Montemitro) (Breu, 2017)	svm	🔊	Glo; Sp	deu;ita	592	URL
Na-našu (San Felice) (Breu, 2017)	svm	🔊	Glo; Sp	deu;ita	628	URL
Stilven (Jaber et al., 2011)	vec	📕	POS	eng	1,450	URL
NLLB-MD (NLLB Team et al., 2022)	fur	📰 👍 ℹ	—	*multi*	8,809	URL
NormLigurian* (Lusito et al., 2023)	lij	📕 W	—	—	6,723	URL
FLORES-200 (NLLB Team et al., 2022)	*multi*$_{(6)}$	W	—	*multi*	12,054	URL
NLLB-Seed (NLLB Team et al., 2022)	*multi*$_{(6)}$	W	—	*multi*	37,159	URL
ITDI (Aepli et al., 2022)	*multi*$_{(11)}$	📕 🌐	—	—	17,886	URL
Stilven (Delmonte et al., 2009)	vec	📕 🪄 📰 🌐	—	eng	9,027	URL

Table 2: Curated corpora for language varieties of Italy used in NLP research (annotated: *top*; unannotated: *bottom*). **Corpora**. Citation and corpus name (*: arbitrary name). **Ids**. ISO 639-3 codes, wherever available; if not, we use an arbitrary designator (*italicized*, cf. Table 1). **Genres**. 📕: narratives, fiction, magazines, novels, children stories; ♫: poems, cantos; 🪄: grammar examples, textbooks; 🔊: transcribed speech or field recordings; 📰: news articles; W: Wikipedia articles; ☁: Bible chapters; 🌐: quotes or proverbs from the Internet; 👍: chat messages; ℹ: non-fiction (incl. health reports). **Annotations**. Morph: morphosyntactic tagging; POS: part-of-speech tagging; Dep: dependency parsing; Glo: glossing; Norm: orthographic normalization; Slot: slot detection; Intent: intent detection; Sp: speech-related information (e.g., pseudo-phones, silences, etc.). **Parallels**. Language(s) of parallel data, if available (*std*: standard orthography; *multi*: many languages). **Sizes**. Number of sentences. **Data**. Link to the publicly available dataset. Notes: *multi*$_{(6)}$ includes fur, lij, lmo, scn, srd and vec; *multi*$_{(11)}$ additionally includes *eml* (Emilian-Romagnol: egl and rgn), lld, pms, nap and *roa-tar* (Tarantino, part of nap).

varieties, text genres, annotation types (if any), languages of parallel data (if applicable), and dataset size (Table 2). Moreover, we inspected Wikisource, Project Gutenberg,[12] UDHR,[13] and raw corpora typically used in multilingual research for the presence of Italy's varieties (Table 3).

Curated corpora for language varieties of Italy (Table 2) greatly vary in terms of objectives, from language documentation (Boito et al., 2018; Breu, 2017) to supporting multilingual information access (Aepli et al., 2023; NLLB Team et al., 2022; van der Goot et al., 2021b). They cover a handful of language varieties (and variants, cf. Section 3.1), are sparse in terms of text genres and annotation types, and are generally small in size. But *(for what) is this a problem?* Both scarcity and sparsity of written content are a challenge to researchers embracing a machine-centric view, who may be tempted to uniformly scale current language technologies to these varieties by creating or crowd-sourcing new written corpora with annotations for a variety of tasks, design "data-efficient" or zero-shot methods to bridge the data scarcity gap, or just build upon raw corpora (Table 3) such as web-crawled text collections regardless of how representative the content and subsequent technologies are of language varieties and speech communities (Section 4.2). Language technologists should here take a step back and thoughtfully reflect on why there is a lack of machine-readable written resources for Italy's varieties and whether this is relevant to the target speech communities. By detaching from the machine-centric view of technology and engaging with speakers of local language varieties, one can realize that most languages and dialects of Italy are primarily oral, have different aspirations for written content and text-based technologies, vary in prospects according to the linguistic and socio-political contexts in which they are embedded, and serve different functions than standardized languages (Section 4.3). Indeed, with the exception of a few language varieties that benefit

[12] https://www.gutenberg.org/.

[13] https://unicode.org/udhr/.

Corpus	#Lang	aae	cim	egl	frp	fur	gem-sty	gem-toi	grk-cal	grk-gri	itk	lij	lld	lmo	mhn	nap	pms	rgn	roa-alc	roa-fae	roa-gar	roa-gis	roa-via	scn	sdc	sdn	sla-res	srd	svm	vec	wae	*eml*	*roa-tar*	*roa-sam*
Project Gutenberg	67					✔										✔																		
Univ. Decl. of Human Rights	529				✔	✔						✔	✔															✔		✔				✔
Wikipedia pages	320	⌛			✔	✔					⌛	✔	✔	✔		✔	✔	⌛						✔	⌛			✔		✔		✔	✔	
Wikisource pages	249					✔						✔	✔	✔		✔	✔							✔				✔		✔				
GNOME v1	187				✔	✔																												
Mozilla-I10n v1	197				✔	✔						✔												✔				✔		✔				
QED v2.0a	225												✔											✔				✔						
sardware v1	2																											✔						
Tatoeba v2023-04-12	397			✔		✔						✔	✔	✔		✔	✔							✔				✔		✔				
Ubuntu v14.10	244				✔	✔						✔	✔			✔	✔											✔		✔	✔			
CCAligned (El-Kishky et al., 2020)	137																											✔						
CCNet (Wenzek et al., 2020)	130													✔			✔																	
OSCAR 22.01 (Abadji et al., 2022)	151													✔			✔							✔								✔		
WikiMatrix (Schwenk et al., 2021)	96													✔										✔										
XLEnt v1.1 (El-Kishky et al., 2021)	120													✔																				

Table 3: Raw corpora including at least one endangered language variety of Italy. *Top*: resources with moderate verification (e.g., open collaboration projects, digitized books, and translations of international documents); *Middle*: crowd-sourced resources (i.e., educational translations and localization files from open-source software projects); *Bottom*: web-crawled resources (i.e., corpora that originate from automated web crawling and are used in multilingual NLP research). **Corpora**. Corpus name, version, and citation, if available. **#Langs**. Number of languages in the corpus. **Presence of varieties**. ✔: present; ⌛: upcoming (in the Wikimedia Incubator as of 2023-09-10). Notes: *eml* (Emilian-Romagnol) is still in use on some sources and can include either `egl` or `rgn`; *roa-tar* (Tarantino) and *roa-sam* (Sammarinese) are typically considered part of `nap` and `rgn`, respectively.

from protection, economic incentives, or co-official status with Italian, and for which a written form is used or envisioned for official purposes (Section 5), written data is likely to remain scarce.

4.2 Little Attention to Representativeness

Another assumption of the machine-centric NLP approach is that, if there is any text collection for a given language variety, it is homogeneous, representative of the community of speakers, and free of noise and boilerplate content, and therefore can be directly used for representing that language variety in language technology. However, unlike the case of standardized languages, most text collections for endangered languages naturally include content in multiple variants, freely written following no consistent or widely established orthography (e.g., Lombard Wikipedia (Miola, 2017)), or comprise a large amount of wrong language and non-linguistic materials (Kreutzer et al., 2022). Nevertheless, those resources are typically taken monolithically regardless of their actual content. In this section, we take Wikipedia and multilingual web-crawled corpora as case studies of mainstream text collections which are used in current NLP regardless of their representativeness of language varieties and speech communities.

Wikipedia is by far the most widely used resource in NLP when it comes to the so-called *under-resourced* languages. It currently comprises content in 320 languages (as of 2023-09-10), of which 10 are endangered varieties of Italy. It additionally includes two more Wikipedias (i.e., `eml`, `roa-tar`) with deprecated or arbitrary language codes. Despite the role of Wikipedia on preserving knowledge even in lesser-used languages, the written content for most endangered varieties has to be taken carefully with regards to the varied guidelines among projects and the potential presence of fictitious and culturally-biased content. For instance, the Lombard Wikipedia leaves users freedom with respect to orthography and local variants (provided that they indicate these on the article page) (Miola, 2017), whereas the written content on the Piedmontese edition of Wikipedia does not match any variety actually spoken (Miola, 2013). A varied use of orthography and local variants can be observed in other Wikipedia editions for Italy's varieties, such as the Ligurian Wikipedia (Lusito and Maillard, 2021). More broadly, the content of small Wikipedias typically comprises translations of pages from larger editions (e.g., English) rather than including original content tied to speakers' identity (Gobbo and Miola, 2016). Besides objectivity, this has the effect to homogenize cultures and perspectives (Callahan and Herring, 2011).

Near-duplicate articles are also common in Wikipedia editions for Italy's varieties. For instance, the Venetian edition of Wikipedia (∼69K pages) contains placeholder content for years from 1 BC to 999 BC (1K pages) and for most of the days of the year, as well as template articles for many municipalities and provinces around the world. This suggests that a relevant portion of the encyclopedia could be generated by bots, and thus that Wikipedia texts for Italy's language varieties not only reflect a rather artificial use of language—what we tentatively call *wikivariety*—but also that the actual content is less than one might think.

Lastly, while `eml` has been deprecated more than 14 years ago[14] in favour of `egl` and `rgn` as separate ethnolinguistic entities (Maiden and Parry, 1997), it is still in use on Wikipedia. Most `eml` pages indicate the specific variety at the top of the article, but this is rarely considered in NLP, where whole Wikipedia editions are taken as monolithic entities for training language models.

The presence of Italy's varieties on Wikipedia has an impact on the creation of web-crawled datasets. It is not surprising that multilingual corpora that include those varieties are the ones that rely on fastText LangID (Joulin et al., 2017), a language identification model that currently includes a handful of Italy's language varieties and whose training material is mostly taken from Wikipedia.

Following Kreutzer et al. (2022), who have recently highlighted systematic issues with web-crawled dataset portions for ''*low-resource* languages'', we manually audit the content of crawled corpora which include Italy's varieties (cf. Table 4) and are easily accessible. The resulting datasets are CCAligned (El-Kishky et al., 2020), WikiMatrix (Schwenk et al., 2021) (parallel), and OSCAR (Abadji et al., 2022) (monolingual).[15]

For each corpus, a native speaker of each included language variety was asked to label a random sample of 50 texts (or parallel texts, in CCAligned and WikiMatrix) according to the labeling scheme and guidelines presented in Kreutzer et al. (2022).[16] Possible labels are c_{nat} (correct, natural), c_{sho} (correct, short), c_{boi} (correct, boilerplate), w_{tra} (wrong translation – applicable to parallel corpora only), w_{lan} (wrong language), and w_{nlg} (wrong, not language). For `lmo` and `scn` in OSCAR, the total instances are less than 50, and thus all of them have been audited. Compared to Kreutzer et al. (2022), we audit data from the latest OSCAR version (22.01), whereas for CCAligned and WikiMatrix we contribute to new language pairs (i.e., en-`srd` and it-`scn`, respectively) and report their results for en-`lmo` on WikiMatrix since we use the same data release and labeling scheme. To go beyond the approach of viewing language varieties as monoliths, native speakers were also asked to mark instances whose variants are hard to categorize because they exhibit traits of continuity with multiple varieties.[17]

	CCA	OSCAR			WIKIMATRIX	
	`srd`	`lmo`	`pms`	`scn`	`lmo`	`scn`
#texts	395	2	698	2	44K	33K
%audit	12.7	100.0	7.2	100.0	<0.1	<0.1
%wiki	18.0	100.0	96.0	100.0	100.0	100.0
c_{nat}	2.0	0.0	100.0	50.0	11.8	16.0
c_{sho}	0.0	0.0	0.0	0.0	0.0	0.0
c_{boi}	30.0	50.0	0.0	50.0	1.0	0.0
w_{tra}	28.0	–	–	–	81.4	78.0
w_{lan}	30.0	50.0	0.0	0.0	4.9	6.0
w_{nlg}	10.0	0.0	0.0	0.0	1.0	0.0
c_{tot}	**32.0**	**50.0**	**100.0**	**100.0**	**12.8**	**16.0**

Table 4: Results (%) of the audit for Italy's language varieties on web-crawled multilingual corpora. The ratio of ''correct'' samples (c_{tot}) is boldfaced.

We present the results of the audit in Table 4. For each corpus and language variety, we also report the number of texts and the percentage of samples we audited. Moreover, we indicate the percentage of Wikipedia content on audited subsets. OSCAR is the corpus with the highest

[14] `https://iso639-3.sil.org/request/2008-040`.

[15] We do not include XLEnt (El-Kishky et al., 2021) since it comprises cross-lingual named entities rather than texts.

[16] In some cases this was not even necessary (e.g., `lmo` in OSCAR, with just 2 instances with unambiguous sources—namely, the `lmo` and `eml` editions of Wikipedia).

[17] Annotations are available in our repository: `https://github.com/varietiesoftheboot/`.

ratio of ‘‘correct’’[18] content (from 50% to 100%); however, very few instances are included in most subsets (e.g., 2 for lmo and scn), and thus results have to be taken with a grain of salt. On the contrary, the previous OSCAR version had more instances, including additional language varieties (Abadji et al., 2022), but the actual linguistic content for most of those was dramatically low, e.g., 0.0% in-language samples for nap (Kreutzer et al., 2022). Interestingly, the sample marked as ‘‘wrong language’’ in the lmo subset comes from the eml Wikipedia edition where it is labeled as *Piacentino*, a variant of egl which exhibits traits of continuity with lmo. This suggests that discretizing variants into bounded languages is rather limiting since they lie on a continuum.

Regarding parallel corpora, most of the content for srd on CCAligned is in another language (30%), a wrong translation (28%), or do not even contain linguistic content (10%). Among the 32% ‘‘correct’’ samples, just one instance (2%) has a clean content. The remaining 30% contain website headers, footers, and other boilerplate content. The situation is even worse on WikiMatrix: While parallel texts are cleaner than CCAligned, most pairs are not translations of each other (81.4% on lmo, 78.0% on scn). The ratio of ‘‘correct’’ content is thus quite low, ranging from 12.8% to 16.0%.

Overall, aside from the domain-specific WikiMatrix, we observe that most of the in-language material for Italy’s language varieties comes from Wikipedia articles. This suggests that content that is not already included in other resources is rarely captured, both because language identifiers trained on Wikipedia are likely to leave nothing but *wikivarieties*, and most importantly because Italy’s varieties are rarely written down, and if so, they are mostly code-switched with a co-territorial ‘‘high-prestige’’ standardized language with vehicular functions, e.g., Italian (Section 4.3).

To conclude, we argue that viewing Italy’s varieties as a *data commodity* for machine learning purposes without asking whether the linguistic content is representative of speech communities disregards the nature of language varieties and ignores their speakers. We encourage researchers *to care* about the varieties they work with and responsibly engage with speech communities (Section 5).

4.3 Uniform Functions, Contexts, and Needs

The strongest assumption of the machine-centric approach is arguably to consider the diverse functions, contexts, and needs of language varieties as homogeneous—and typically in the image of *high-resource* standardized languages, e.g., Italian or English. This practice has the effect to reduce language varieties to mere linguistic codes that are dissociated from their distinctive situations.

By looking at the contemporary sociolinguistic context of Italy, most local language varieties exist in a situation of *dilalìa* (Berruto, 1987) with the national language. While Italian serves as the ‘‘high-prestige’’ vehicular language (Fishman, 2001), and it is therefore the language used in all formal settings (i.e., from education to administration), Italy’s languages and dialects are primarily confined to spoken, informal situations (e.g., family, local participation), and Italian functionally overlaps with them in those informal domains—making the situation different from the rigidly compartmentalized *diglossia* (Avolio, 2009). Exceptions are language varieties within territories in which bilingualism is officially granted by national laws, i.e., those of the German minority in the South Tyrol province (northern Italy), the French minority in the Aosta Valley (northwest Italy), and the Slovenian minority in some municipalities of the Friuli-Venezia Giulia region (northeast Italy). In those cases, local language varieties typically enjoy the same standing of the national language and are used (or are aimed to be used) to serve ‘‘high-prestige’’ functions. This functional differentiation should be the starting point for language technologists to reflect on the (often considered homogeneous) utility of text-based language technologies across language varieties.

The socio-political contexts in which language varieties are situated have an impact on language vitality prospects and community aspirations, too. For instance, some language varieties and their culture are protected by the Italian Law 482/1999 (1999),[19] albeit safeguarding measures differ on how they are locally implemented. Moreover,

[18] As in Kreutzer et al. (2022), ‘‘correct’’ indicates that the written variant in the sample is clearly part of a language variety. It does not aim to determine a ‘‘correct form of writing’’.

[19] Those are the ones of the Albanian, Catalan, Germanic, Greek, Slovenian, Croatian, French, Francoprovençal, Friulian, Ladin, Occitan, and Sardinian speech communities.

some of them also benefit from recognition and safeguard by regional laws,[20] or are even locally co-official (i.e., German and Ladin in the South Tyrol province and French in the Aosta Valley). Finally, some language varieties are solely recognized or promoted locally, or both.[21] These diverse situations must be attentively considered, and engaging with local communities would allow the researcher to deeply understand how this affects their ambitions and needs (Section 5).

As regards written use, although some language varieties of Italy have a notable literary tradition (Avolio, 2009) (e.g., works in Venetian by C. Goldoni [18th century] and in Neapolitan by G. Basile [16th century], *inter alia*), we stress that they are nowadays primarily used in spoken, informal settings, and most of them have no standardized written form. Even if official orthography standards exist for some varieties, these are often unknown to speakers themselves. Indeed, in our experience speakers write ''the way words sound'' in their local variants, using just the available characters in their keyboards. Normalizing user-generated texts to a ''standard'' form (e.g., Baldwin et al., 2015; van der Goot et al., 2020, 2021a) has proven useful for NLP purposes, but it inevitably erases the naturally occurring sociolinguistic variation (Nguyen et al., 2021), homogenizing all variants of a language variety and imposing a ''correct'' form of writing.

But *how often do speakers write in their own variety?* With the exception of restricted communities on social media and few dedicated websites, writing in some of Italy's language varieties is rather uncommon. Code-switching—the alternation of different language varieties in a single discourse—is instead a more widespread practice in Italy (Cerruti and Regis, 2005), where Standard Italian—or any co-territorial ''high-prestige'' language in border areas—is mixed with both Italy's varieties and regional Italian. This brings into question the utility of sentence- and document-level language identification tools supporting Italy's language varieties.

5 Towards a *Speaker-centric* Approach

The assumptions and shortcomings discussed in the preceding sections make evident that the current machine-centric approach neither respects nor represents language varieties of Italy and their speakers. Ultimately, language technology should serve speech communities and their language varieties, and not the other way round. We need to identify new ways of working that are centered on speech communities and their varieties—what we refer to as *speaker-centric* approach. In this section, we provide recommendations and opportunities towards speaker-centric work that foresees active engagement with speech communities.

Becoming Aware of Local History and Diverse Attitudes Before starting to engage with speech communities, it is advisable to become aware that local language varieties may be perceived very differently by their own speakers. Local languages and dialects of Italy have been historically subjected to prejudices and censorship. This culminated with the *Italianization* policy implemented by the fascist regime in 1923–1942 whereby ''[local language varieties] were banned in the most absolute way [...] even when playing with classmates'' (Camilleri and De Mauro, 2014), teaching in languages other than Italian was abolished, and foreign toponyms and surnames were changed to Italian-sounding forms. Among other things, this contributed over the next half century to the continued view of local language varieties as a ''synonym of ignorance and lack of integration'' (D'Agostino, 2015). Recent years have instead witnessed an overall change in attitude on the matter, especially by the young, for whom local varieties are rather rediscovered as an additional expressive resource in their communicative repertoire (Berruto, 2006). It is therefore necessary to realize in advance that—even within the same community—we may encounter speakers with diverse sensitivities and motivations, and that those may also be influenced by political parties that leverage language varieties for independence purposes. We need to remember that speech communities do not have a ''single voice'' (Bird, 2020)

[20] Arbëreshë Albanian in Apulia and Calabria regions; Algherese Catalan, Gallurese, Sardinian, and Sassarese in Sardinia; German in the Walser-speaking Valle del Lys (Aosta Valley); Cimbrian, Ladin, and Mòcheno in Trentino; Calabrian Greek and Occitan (i.e., Vivaro-Alpine Gardiol) in Calabria; Francoprovençal (i.e., Faetar) and Griko in Apulia.

[21] Recognized: Lombard, Piedmontese, and Sicilian in Lombardy, Piedmont, and Sicily, respectively; Promoted: Friulian and Slovenian in Friuli-Venezia Giulia, and Francoprovençal, French, Occitan, and Walser in Piedmont; Both: Venetian in Veneto and Ligurian *Tabarchino* in Sardinia.

and that language ideologies and practices may change and be embraced differently over time (e.g., Griko in Pellegrino, 2021).

Engaging with Local Communities Building relationships with speech communities (Liu et al., 2022; Schwartz, 2022; Bird, 2020) is pivotal for speaker-centric work. It allows researchers not only to get a better sense of local communities' attitudes and aspirations, and understand the individual linguistic and socio-political contexts at the micro-level, but also to learn about local agendas to support language vitality. However, the engagement should not be for the sole benefit of the researcher, but rather based on equity, reciprocity, and respect (Bird, 2020). From here naturally comes mutual trust, deep understanding of community needs, and thus opportunities for locally meaningful language technology applications—that may range from online dictionaries, to computer-assisted language education, to multilingual information access, depending on the individual situation.[22] In the context of Italy, it is important to note that the engagement process and involved actors may differ across communities. For instance, very small communities in which language varieties are mostly spoken by elders (e.g., Cimbrian, Calabrian Greek; cf. Table 1) are represented by a number of cultural institutes that occasionally promote initiatives on language and culture. Participating to local events, understanding customs and traditions, and ask curiosity-driven questions is probably the only way to start building meaningful bonds in this space.[23] Instead, larger speech communities of non-officially recognized varieties (e.g., Neapolitan, Venetian; cf. Table 1) are often supported by politically-polarized bodies, but language varieties are spoken even by younger generations (ISTAT, 2017). It is advisable here to engage with individuals with diverse backgrounds and demographic characteristics. Given the number of speakers of those varieties, if casual relationships are not already in place, bonds can be easily established in the most diverse environments, including academia. Once a collaboration space between communities and NLP researchers is found, the involvement of speech communities must not end. In the speaker-centric approach, communities are involved at multiple stages of the design process, inspired by participatory design methods (Caselli et al., 2021). External language technologists need to recall that they work with others' data for supporting vitality of others' language varieties, and that only speech communities can reliably judge the usefulness and representativeness of a given technological artifact, both during and after the process. About representativeness, it is important to acknowledge that language and culture are inseparable, and that current NLP is not culturally sensitive (Hershcovich et al., 2022). Shared knowledge may differ from place to place, and this indeed shapes language. It would not be surprising if a machine translation system for Cimbrian—assuming that this is actually needed—homogenized ‘‘snow’’ to a single word, regardless of the many names it gets in Cimbrian highlands according to seasons and conditions (Rigoni Stern, 1998). Broadly, this is a motivation for NLP to start shifting from the traditional, monocultural view of language to a more inclusive, culturally-aware language technology. Moreover, it opens opportunities at the intersection of participatory design and NLP, e.g., new evaluation methods based on continuous communities' feedback.

Building a Community Responsibly supporting the vitality of language varieties of Italy by adopting a speaker-centric approach could be a difficult process to initiate. Moreover, in pursing this goal we may find it valuable to build concrete relationships with other stakeholders, exchange local knowledge and experience, and establish collaborations across speech communities (e.g., those sharing similar aspirations or which language varieties are closely related) and researchers from different academic disciplines (e.g., NLP, linguistics, anthropology). To ease this process, we initiated Varieties of the Boot,[24] a community aimed at responsibly supporting the vitality of language varieties of Italy by *i)* offering guidance on the speaker-centric approach to individuals interested in engaging in this space, *ii)* fostering discussion on practices that have been adopted in the past in diverse environments, lessons learnt, and mistakes to be avoided, and *iii)* encouraging

[22] Indeed, it would be simplistic in the context of this paper to suggest specific language technologies for each variety.

[23] To encourage researchers' awareness and participation in these contexts, we provide a collection of language and culture institutes and related entities in our repository: `https://github.com/varietiesoftheboot/`.

[24] `https://varietiesoftheboot.github.io/`.

participatory work between diverse speech communities, cultural institutes, and fields of study. Finally, *iv)* the community intends to serve as a reference point for actively raising awareness among the Italian community at large and external researchers about the often overlooked linguistic heritage of Italy. Practically, this may not only include scientific events such as thematic workshops, but also local events and communication activities on social media. The community opens valuable opportunities for stakeholders to learn from diverse perspectives, to responsibly engage with speech communities at different places, and to start participatory, interdisciplinary and intercultural collaborations.

Pursuing Alternative Directions There are many opportunities for NLP in neighboring areas. Language technology has traditionally focused on Standard Italian, but in everyday communication Italian speakers are instead used to use their own form of regional Italian (Avolio, 2009) (Section 2.3), i.e., varieties resulting from the geographical differentiation of the standard language. Ultimately, NLP should better represent the actual use of the Italian language. This also opens opportunities to study fairness of current NLP models across regional variants. Moreover, NLP to study language variation and contact at scale (Ramponi and Casula, 2023; Hovy and Purschke, 2018; Donoso and Sánchez, 2017, *inter alia*) can help in documenting how regional Italian varies across space. This can ultimately enrich and complement existing linguistic atlases such as ALI (Bartoli et al., 1995) and AIS (Jaberg et al., 1987). Finally, based on the actual use of Italy's varieties, studying code-switching with a focus on its linguistic and social context (Doğruöz et al., 2021) may contribute to understanding language replacement processes (Cerruti and Regis, 2005).

6 Conclusion

In this work, we present the complex linguistic landscape of Italy, shedding light on the main assumptions and shortcomings of the default, *machine-centric* NLP approach for local language varieties. We advocate for a shift in the paradigm towards *speaker-centric* NLP, and provide recommendations and opportunities for responsible, participatory work aimed to support vitality of language varieties of Italy, designed *with* speech communities, *for* serving speakers and their needs.

Acknowledgments

We would like to thank the action editor and the anonymous reviewers for their insightful and constructive feedback during the review process. We would also like to thank Sara Tonelli and Barbara Plank for their advice on earlier versions of this paper, and the members of the Digital Humanities group at Fondazione Bruno Kessler for the precious conversations and contribution to data auditing. Further, we are grateful to Camilla Amendola for her valuable assistance in designing Figure 1.

References

Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. Towards a cleaner document-oriented multilingual crawled corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4344–4355, Marseille, France. European Language Resources Association.

Ahmed Abdelali, Francisco Guzman, Hassan Sajjad, and Stephan Vogel. 2014. The AMARA corpus: Building parallel language resources for the educational domain. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1856–1862, Reykjavik, Iceland. European Language Resources Association (ELRA).

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.

Noëmi Aepli, Çağrı Çöltekin, Rob Van Der Goot, Tommi Jauhiainen, Mourhaf Kazzaz, Nikola Ljubešić, Kai North, Barbara Plank, Yves Scherrer, and Marcos Zampieri. 2023. Findings of the VarDial evaluation campaign 2023. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial*

2023), pages 251–261, Dubrovnik, Croatia. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.vardial-1.25

Maristella Agosti, Birgit Alber, Giorgio Maria Di Nunzio, Marco Dussin, Stefan Rabanus, and Alessandra Tomaselli. 2012. A curated database for linguistic research: The test case of Cimbrian varieties. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2230–2236, Istanbul, Turkey. European Language Resources Association (ELRA).

Antonios Anastasopoulos, Marika Lekakou, Josep Quer, Eleni Zimianiti, Justin DeBenedetto, and David Chiang. 2018. Part-of-speech tagging on an endangered language: A parallel Griko-Italian resource. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2529–2539, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Manuela Angioni, Franco Tuveri, Maurizio Virdis, Laura Lucia Lai, and Micol Elisa Maltesi. 2018. SardaNet: A linguistic resource for Sardinian language. In *Proceedings of the 9th Global Wordnet Conference*, pages 412–419, Nanyang Technological University (NTU), Singapore. Global Wordnet Association.

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

Timothy Baldwin, Marie Catherine de Marneffe, Bo Han, Young-Bum Kim, Alan Ritter, and Wei Xu. 2015. Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 126–135, Beijing, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/W15-4319

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

Delphine Bernhard, Anne-Laure Ligozat, Myriam Bras, Fanny Martin, Marianne Vergez-Couret, Pascale Erhart, Jean Sibille, Amalia Todirascu, Philippe Boula de Mareüil, and Dominique Huck. 2021. Collecting and annotating corpora for three under-resourced languages of France: Methodological issues. *Language Documentation & Conservation*, 15:316–357.

Delphine Bernhard, Anne-Laure Ligozat, Fanny Martin, Myriam Bras, Pierre Magistry, Marianne Vergez-Couret, Lucie Steiblé, Pascale Erhart, Nabil Hathout, Dominique Huck, Christophe Rey, Philippe Reynés, Sophie Rosset, Jean Sibille, and Thomas Lavergne. 2018. Corpora with part-of-speech annotations for three regional languages of France: Alsatian, Occitan and Picard. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Gaetano Berruto. 1987. Lingua, dialetto, diglossia, dilalìa. *Romania et Slavia Adriatica*, pages 57–81. Buske, Hamburg, Germany.

Gaetano Berruto. 2005. Dialect/standard convergence, mixing, and models of language contact: The case of Italy. *Dialect Change: Convergence and Divergence in European Languages*, pages 81–95. https://doi.org/10.1017/CBO9780511486623.005

Gaetano Berruto. 2006. Quale dialetto per l'Italia del duemila? Aspetti dell'italianizzazione e risorgenze dialettali in Piemonte (e altrove). In *Lingua e Dialetto nell'Italia del Duemila*, pages 101–127. Congedo, Galatina, Italy.

Steven Bird. 2020. Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519, Barcelona, Spain (Online). International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.313

Steven Bird. 2022. Local languages, third spaces, and other high-resource scenarios. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7817–7829, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.539

Marcely Zanon Boito, Antonios Anastasopoulos, Aline Villavicencio, Laurent Besacier, and Marika Lekakou. 2018. A small Griko-Italian speech translation corpus. In *Proceedings of the 6th International Workshop on Spoken Language Technologies for Under-Resourced Languages*, pages 36–41. https://doi.org/10.21437/SLTU.2018-8

Walter Breu. 2017. *Slavische Mikrosprachen im Absoluten Sprachkontakt: Glossierte und Interpretierte Sprachaufnahmen aus Italien, Deutschland, Österreich und Griechenland. Teil I: Moliseslavische Texte aus Acquaviva Collecroce, Montemitro und San Felice del Molise*. Harrassowitz Verlag, Wiesbaden, Germany. https://doi.org/10.2307/j.ctv11sn5zw

Ewa S. Callahan and Susan C. Herring. 2011. Cultural bias in Wikipedia content on famous persons. *Journal of the American Society for Information Science and Technology*, 62(10):1899–1915. https://doi.org/10.1002/asi.21577

Andrea Camilleri and Tullio De Mauro. 2014. *La Lingua Batte dove il Dente Duole*, 3rd edition. I Robinson. Letture. Laterza, Roma-Bari, Italy.

Tommaso Caselli, Roberto Cibin, Costanza Conforti, Enrique Encinas, and Maurizio Teli. 2021. Guiding principles for participatory design-inspired natural language processing. In *Proceedings of the 1st Workshop on NLP for Positive Impact*, pages 27–35, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.nlp4posimpact-1.4

Massimo Cerruti. 2011. Regional varieties of Italian in the linguistic repertoire. *International Journal of the Sociology of Language*, 2011(210):9–28. https://doi.org/10.1515/ijsl.2011.028

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

Aditi Chaudhary, Antonios Anastasopoulos, Zaid Sheikh, and Graham Neubig. 2021. Reducing confusion in active learning for part-of-speech tagging. *Transactions of the Association for Computational Linguistics*, 9:1–16. https://doi.org/10.1162/tacl_a_00350

Costanza Conforti and Alexander Fraser. 2017. Supervised word sense disambiguation for Venetan: A proof-of-concept experiment. In *The Thirtieth International Flairs Conference*, Marco Island, Florida. Association for the Advancement of Artificial Intelligence.

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. https://doi.org/10.18653/v1/2020.acl-main.747

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Arya D. McCarthy, Katharina Kann, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. The CoNLL–SIGMORPHON 2018 shared task: Universal morphological reinflection. In *Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, pages 1–27, Brussels. Association for Computational Linguistics. https://doi.org/10.18653/v1/K18-3001

Mari D'Agostino. 2015. Sociolinguistica dell'italiano contemporaneo. In *L'Italia e le sue Regioni*, volume III. Treccani, Roma, Italy.

Tullio De Mauro. 1963. *Storia Linguistica dell'Italia Unita*. Laterza, Roma-Bari, Italy.

Rodolfo Delmonte, Antonella Bristot, Sara Tonelli, and Emanuele Pianta. 2009. English/Veneto resource poor machine translation with STILVEN. In *Proceedings of ISMTCL*, pages 82–89, Besançon, France. Presses Universitaires de Franche-Comté.

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. https://doi.org/10.18653/v1/N19-1423

A. Seza Doğruöz, Sunayana Sitaram, Barbara E. Bullock, and Almeida Jacqueline Toribio. 2021. A survey of code-switching: Linguistic and social perspectives for language technologies. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1654–1666, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.131

Gonzalo Donoso and David Sánchez. 2017. Dialectometric analysis of language variation in Twitter. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 16–25, Valencia, Spain. Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-1202

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2022. *Ethnologue: Languages of the World*, twenty-fifth edition. SIL International, Dallas, Texas.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. CCAligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5960–5969, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.480

Ahmed El-Kishky, Adithya Renduchintala, James Cross, Francisco Guzmán, and Philipp Koehn. 2021. XLEnt: Mining a large cross-lingual entity dataset with lexical-semantic-phonetic word alignment. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10424–10430, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.814

Tomaž Erjavec. 2017. MULTEXT-East. In *Handbook of Linguistic Annotation*, pages 441–462. Springer, Dordrecht, Netherlands. https://doi.org/10.1007/978-94-024-0881-2_17

Joshua A. Fishman. 2001. Why is it so hard to save a threatened language? In *Can Threatened Languages be Saved?*, pages 1–22. Multilingual Matters, Bristol, Blue Ridge Summit. https://doi.org/10.21832/9781853597060-003

Jennifer-Carmen Frey, Aivars Glaznieks, and Egon W. Stemle. 2016. The DiDi corpus of South Tyrolean CMC data: A multilingual corpus of Facebook texts. In *Proceedings of the Third Italian Conference on Computational Linguistics CLiC-it 2016*, Napoli, Italy. Accademia University Press. https://doi.org/10.4000/books.aaccademia.1782

Gianfranco Fronteddu, Hèctor Alòs i Font, and Francis M. Tyers. 2017. Una eina per a una llengua en procés d'estandardització: El traductor automàtic català-sard. *Linguamática*, 9(2):3–20. https://doi.org/10.21814/lm.9.2.255

Samuel Frontull. 2022. Machine Translation for the Low-resource Ladin of the Val Badia. Master's thesis, University of Innsbruck.

Aivars Glaznieks, Jennifer-Carmen Frey, Maria Stopfner, Lorenzo Zanasi, and Lionel Nicolas. 2022. Leonide: A longitudinal trilingual corpus of young learners of Italian, German and English. *International Journal of Learner Corpus Research*, 8(1):97–120. https://doi.org/10.1075/ijlcr.21004.gla

Federico Gobbo and Emanuele Miola. 2016. Modificare l'immagine linguistica: Esperanto e Piemontese a confronto. In *Représentations Sociales des Langues et Politiques Linguistiques. Déterminismes, Implications, Regards Croisés*, pages 287–304, Roma, Italy. Aracne Editrice.

Rob van der Goot, Alan Ramponi, Tommaso Caselli, Michele Cafagna, and Lorenzo De Mattei. 2020. Norm it! Lexical normalization for Italian and its downstream effects for dependency parsing. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6272–6278, Marseille, France. European Language Resources Association.

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. 2021a. 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. https://doi.org/10.18653/v1/2021.wnut-1.55

Rob van der Goot, Ibrahim Sharaf, Aizhan Imankulova, Ahmet Üstün, Marija Stepanović, Alan Ramponi, Siti Oryza Khairunnisa, Mamoru Komachi, and Barbara Plank. 2021b. From masked language modeling to translation: Non-English auxiliary tasks improve zero-shot spoken language understanding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2479–2497, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.197

Ken Hale, Michael Krauss, Lucille J. Watahomigie, Akira Y. Yamamoto, Colette Craig, LaVerne Masayesva Jeanne, and Nora C. England. 1992. Endangered languages. *Language*, 68(1):1–42. https://doi.org/10.2307/416368

Atticus Harrigan, Aditi Chaudhary, Shruti Rijhwani, Sarah Moeller, Antti Arppe, Alexis Palmer, Ryan Henke, and Daisy Rosenblum, editors. 2023. *Proceedings of the Sixth Workshop on the Use of Computational Methods in the Study of Endangered Languages*. Association for Computational Linguistics, Remote.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.482

Erhard Hinrichs and Steven Krauwer. 2014. The CLARIN research infrastructure: Resources and tools for eHumanities scholars. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1525–1531, Reykjavik, Iceland. European Language Resources Association (ELRA).

Dirk Hovy and Christoph Purschke. 2018. Capturing regional variation with distributed place representations and geographic retrofitting. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4383–4394, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1469

ISTAT. 2017. L'uso della lingua italiana, dei dialetti e di altre lingue in Italia. https://www.istat.it/it/archivio/207961. Accessed: 2023-09-10.

Italian Law 482/1999. 1999. Norme in materia di tutela delle minoranze linguistiche storiche. https://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:legge:1999;482. Accessed: 2023-09-10.

Suhel Jaber, Sara Tonelli, and Rodolfo Delmonte. 2011. Venetan to English machine translation: Issues and possible solutions. In *Proceedings of the 8th International NLPSC Workshop*, pages 69–80, Copenhagen, Denmark. Samfundslitteratur.

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

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.560

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics. https://doi.org/10.18653/v1/E17-2068

Ben King and Steven Abney. 2013. Labeling the languages of words in mixed-language documents using weakly supervised methods. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1110–1119, Atlanta, Georgia. Association for Computational Linguistics.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72. https://doi.org/10.1162/tacl_a_00447

Marika Lekakou, Valeria Baldiserra, and Antonis Anastasopoulos. 2013. Documentation and analysis of an endangered language: Aspects of the grammar of Griko. http://griko.project.uoi.gr/. Accessed: 2023-05-01.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742. https://doi.org/10.1162/tacl_a_00343

Zoey Liu, Crystal Richardson, Richard Hatcher, and Emily Prud'hommeaux. 2022. Not always about you: Prioritizing community needs when developing endangered language technology. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3933–3944, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.272

Stefano Lusito, Edoardo Ferrante, and Jean Maillard. 2023. Text normalization for low-resource languages: The case of Ligurian. In *Proceedings of the Sixth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 98–103, Remote. Association for Computational Linguistics.

Stefano Lusito and Jean Maillard. 2021. A Universal Dependencies corpus for Ligurian. In *Proceedings of the Fifth Workshop on Universal Dependencies (UDW, SyntaxFest 2021)*, pages 121–128, Sofia, Bulgaria. Association for Computational Linguistics.

Manuel Mager, Arturo Oncevay, Annette Rios, Ivan Vladimir Meza Ruiz, Alexis Palmer, Graham Neubig, and Katharina Kann, editors. 2021. *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*. Association for Computational Linguistics, Online.

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

Dan Malkin, Tomasz Limisiewicz, and Gabriel Stanovsky. 2022. A balanced data approach for evaluating cross-lingual transfer: Mapping the linguistic blood bank. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4903–4915, Seattle, United States. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.361

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308. https://doi.org/10.1162/coli_a_00402

Arya D. McCarthy, Ekaterina Vylomova, Shijie Wu, Chaitanya Malaviya, Lawrence Wolf-Sonkin, Garrett Nicolai, Christo Kirov, Miikka

Silfverberg, Sabrina J. Mielke, Jeffrey Heinz, Ryan Cotterell, and Mans Hulden. 2019. The SIGMORPHON 2019 shared task: Morphological analysis in context and cross-lingual transfer for inflection. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–244, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-4226

Emanuele Miola. 2013. A sociolinguistic account of WikiPiedmontese and WikiLombard / Eine soziolinguistische auswertung von Wiki-Piedmontesisch und Wiki-Lombardisch / Etude sociolinguistique des Wikipédias en Piémontais et en Lombard. *Sociolinguistica*, 27(1):116–131. https://doi.org/10.1515/soci.2013.27.1.116

Emanuele Miola. 2017. Dalla parola alla scrittura: Il caso di Emiliano, Veneto e Siciliano. *Quaderni di Linguistica (La Scrittura all'Ombra della Parola)*, 5:59–72.

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

Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors. 2022. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland.

Dong Nguyen, Laura Rosseel, and Jack Grieve. 2021. On learning and representing social meaning in NLP: A sociolinguistic perspective. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 603–612, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.50

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No Language Left Behind: Scaling human-centered machine translation. *CoRR*, arXiv:2207.04672v3. https://doi.org/10.48550/arXiv.2207.04672

Atul Kr. Ojha, Sina Ahmadi, Chao-Hong Liu, and John P. McCrae, editors. 2022. *Proceedings of the Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia within the 13th Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France.

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

Manuela Pellegrino. 2021. *Greek Language, Italian Landscape: Griko and the Re-storying of a Linguistic Minority*. Hellenic Studies Series. Harvard University, Center for Hellenic Studies, Washington, DC, USA.

Alan Ramponi and Camilla Casula. 2023. DiatopIt: A corpus of social media posts for the study of diatopic language variation in Italy. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 187–199, Dubrovnik, Croatia. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.vardial-1.19

Mario Rigoni Stern. 1998. *Sentieri sotto la Neve*. Supercoralli. Einaudi, Torino, Italy.

Yves Scherrer, Tommi Jauhiainen, Nikola Ljubešić, Preslav Nakov, Jörg Tiedemann, and Marcos Zampieri, editors. 2023. Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023). Association for Computational Linguistics, Dubrovnik, Croatia.

Lane Schwartz. 2022. Primum Non Nocere: Before working with Indigenous data, the ACL must confront ongoing colonialism. In

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 724–731, Dublin, Ireland. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2022.acl-short.82`

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2021.eacl-main.115`

Gary Simons and Steven Bird. 2003. The Open Language Archives Community: An infrastructure for distributed archiving of language resources. *Literary and Linguistic Computing*, 18(2):117–128. `https://doi.org/10.1093/llc/18.2.117`

Sarah G. Thomason. 2015. *Endangered Languages: An Introduction*. Cambridge Textbooks in Linguistics. Cambridge University Press, Cambridge, England. `https://doi.org/10.1017/CBO9781139033817`

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Sara Tonelli, Emanuele Pianta, Rodolfo Delmonte, and Michele Brunelli. 2010. VenPro: A morphological analyzer for Venetan. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Francis M. Tyers, Hèctor Alòs i Font, Gianfranco Fronteddu, and Adrià Martín-Mor. 2017. Rule-based machine translation for the Italian-Sardinian language pair. *The Prague Bulletin of Mathematical Linguistics*, 108:221–232. `https://doi.org/10.1515/pralin-2017-0022`

Ekaterina Vylomova, Jennifer White, Elizabeth Salesky, Sabrina J. Mielke, Shijie Wu, Edoardo Maria Ponti, Rowan Hall Maudslay, Ran Zmigrod, Josef Valvoda, Svetlana Toldova, Francis Tyers, Elena Klyachko, Ilya Yegorov, Natalia Krizhanovsky, Paula Czarnowska, Irene Nikkarinen, Andrew Krizhanovsky, Tiago Pimentel, Lucas Torroba Hennigen, Christo Kirov, Garrett Nicolai, Adina Williams, Antonios Anastasopoulos, Hilaria Cruz, Eleanor Chodroff, Ryan Cotterell, Miikka Silfverberg, and Mans Hulden. 2020. SIGMORPHON 2020 shared task 0: Typologically diverse morphological inflection. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 1–39, Online. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2020.sigmorphon-1.1`

Takashi Wada, Tomoharu Iwata, Yuji Matsumoto, Timothy Baldwin, and Jey Han Lau. 2021. Learning contextualised cross-lingual word embeddings and alignments for extremely low-resource languages using parallel corpora. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 16–31, Punta Cana, Dominican Republic. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2021.mrl-1.2`

Eryk Wdowiak. 2022. A recipe for low-resource NMT. In *Intelligent Computing (SAI 2022)*, pages 739–746, Cham, Switzerland. Springer International Publishing. `https://doi.org/10.1007/978-3-031-10464-0_50`

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.