EMNLP 2023 – *Singapore, Dec 6-10, 2023*

Language Varieties of Italy

Technology Challenges and Opportunities

TACL – *Transactions of the Association for Computational Linguistics*

Alan Ramponi

Fondazione Bruno Kessler, Italy

The *linguistic context* of Italy

> *"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)

Local language varieties

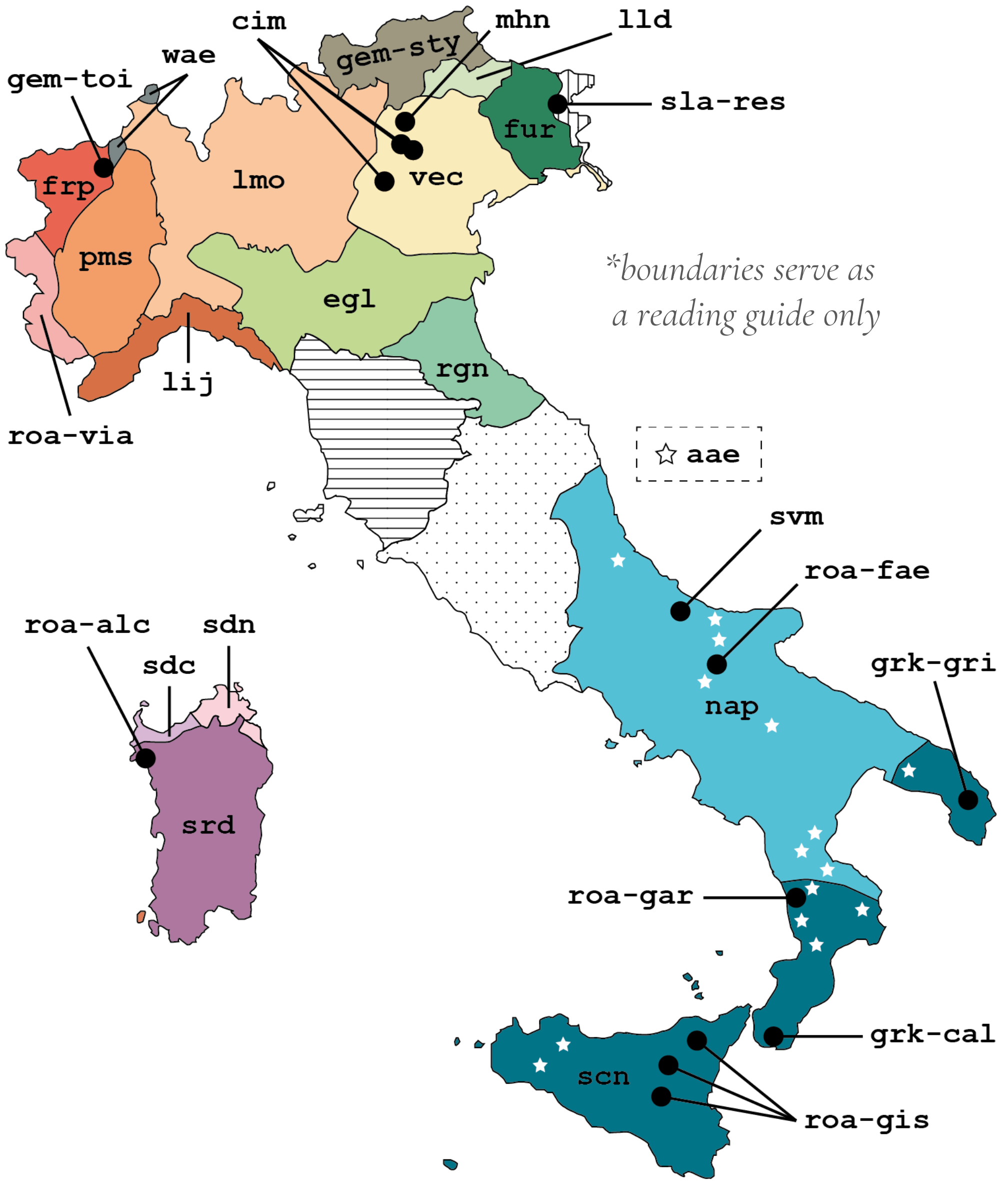
Primarily used in **spoken contexts**

- *Romance* varieties: most are **"sisters" of Italian**
- *Germanic*, *Albanian*, *Hellenic*, and *Slavic* ones

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

○ *vulnerable* ◉ *definitely endangered* ● *severely endangered* *UNESCO (Moseley, 2010)*

Moseley, 2010. "Atlas of the World's Languages in Danger". UNESCO Publishing.



**boundaries serve as a reading guide only*

The default *machine-centric* approach

Strong emphasis on *data scarcity*

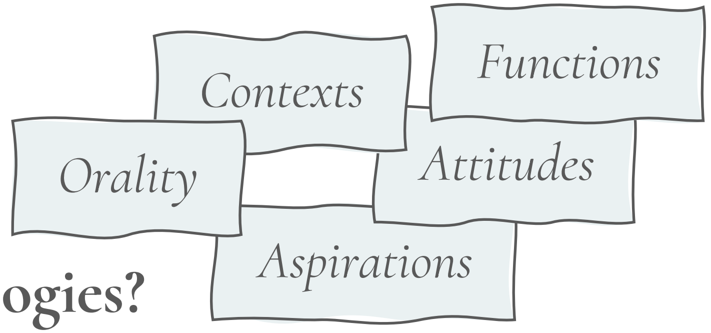
"Under-resourcedness" (⚠ in terms of *machine-readable*, *written* data)

⚠ Common take: need for more resources or computational means to bridge the gap

- Counting: decoupling *unique situations* from *volume of resources* (Joshi et al., 2020)

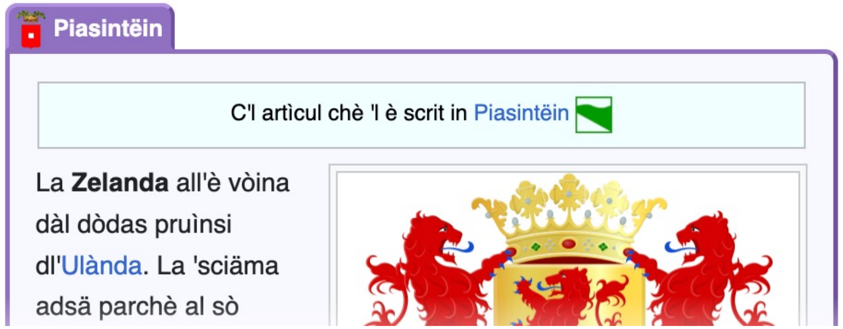
Just a step back…

- **Why written resources are scarce (and sparse)?**
- **Do target communities really need text-based technologies?**



Joshi et al., 2020. "*The State and Fate of Linguistic Diversity and Inclusion in the NLP World*". *ACL*.

On *representativeness*: Wikipedia



Piasintëin

C'l articul chè 'l è scrit in Piasintëin

La **Zelanda** all'è vòina dàl dòdas pruìnsi dl'Ulànda. La 'sciäma adsä parchè al sò

Mainstream resource (320 editions, *10+ Italy's varieties*)

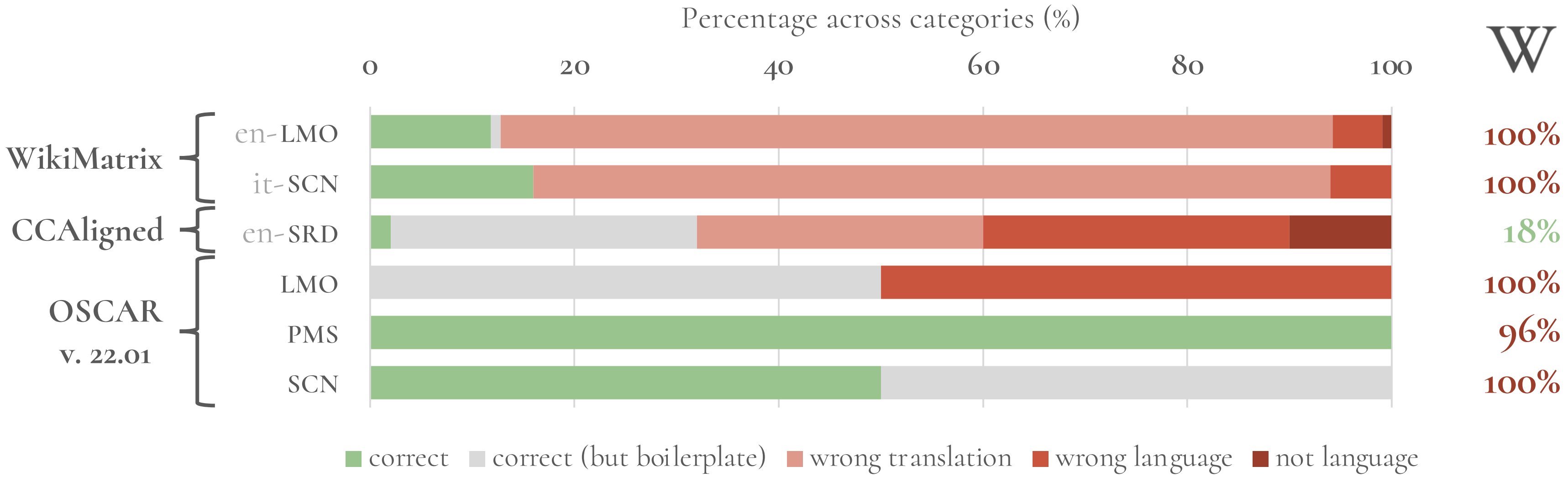
⚠ *Typically taken monolithically* regardless of their actual content and metadata

- **Different editions, different guidelines**: *orthographies* and *local variants*
- **Content not tied to speakers' identity**: *homogeneization* of cultures/perspectives
- **Artificial varieties (*wikivarieties*)**: no *local themes & lexicon* (e.g., *objects*, *professions*)
- **Partially bot-generated**: many *placeholder* pages (e.g., years, municipalities, …)

On *representativeness*: web-crawled corpora

Manual audit of crawled corpora which include Italy's varieties

- Following the labeling scheme and guidelines presented in Kreutzer et al. (2022)

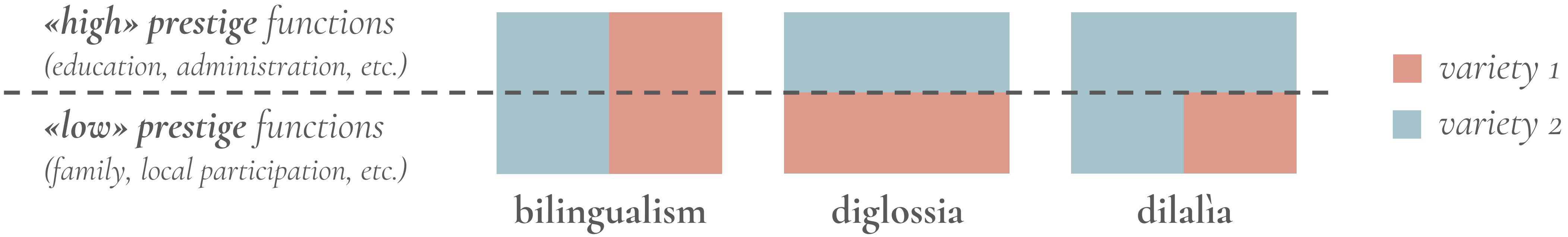


⚠ lang. identification tools trained on Wikipedia leave mostly ***wikivarieties***

***Kreutzer et al., 2022**. "Quality at a Glance: An Audit of Web-crawled Multilingual Datasets". TACL.*

Uniform *functions* and *contexts* – and *needs*

Functional differentiation: are language varieties all the same?



«high» prestige functions
(education, administration, etc.)

«low» prestige functions
(family, local participation, etc.)

bilingualism **diglossia** **dilalìa**

🟥 *variety 1*
🟦 *variety 2*

Sociopolitical contexts: some are protected by the Italian Law 482/1999, others by regional laws, others locally co-official, others promoted locally

Actual use: primarily oral, code-switched, written "the way words sound"

Towards a *speaker-centric* approach

Becoming aware of *history* & *attitutes*

Language varieties are **perceived differently** by their speakers

- Historically subjected to ***prejudices*** & ***censorship*** (e.g., *Fascist Italianization*)
- Then seen as synonym of ***ignorance*** & ***lack of integration*** (D'Agostino, 2015)
- Leveraged by political parties for ***independence purposes*** (esp. *northern Italy*)
- Rediscovered as ***additional expressive resource*** in comm. reportoire (Berruto, 2006)

⚠ Speech communities have many voices that may change over time

D'Agostino, 2015. *"Sociolinguistica dell'Italiano Contemporaneo". L'Italia e le sue Regioni, vol III., Treccani.*
Berruto, 2006. *"Quale Dialetto per l'Italia del Duemila?". Lingua e Dialetto nell'Italia del Duemila.*

Engaging with *local communities*

Understand the *cultural*, *linguistic & socio-political* context

- **Learn** about *local agendas* to support language vitality (→ *locally-meaningful work*)
  - E.g., culture preservation, language learning, intergenerational transmission
- **Engage** with local communities following *equity*, *reciprocity*, and *respect* (Bird, 2020)
- **Involve** speakers at all design stages (e.g., *participatory work*, Caselli et al. (2021))

⚠ *Engagement process & actors* may differ across communities/contexts

- *Cultural institutes* promoting initiatives on language/culture *vs* *individuals*

“**Language and Culture Institutes in Italy**”. *https://github.com/varietiesoftheboot/language-and-culture-institutes*.

***Bird*, 2020**. “*Decolonising Speech and Language Technology*”. *COLING.*

***Caselli et al.*, 2021**. “*Guiding Principles for Participatory Design-inspired Natural Language Processing*”. *NLP4PI@ACL* https://varietiesoftheboot.github.io

Initiating *Varieties of the Boot*

*A local, multidisciplinary community aimed at studying and supporting the vitality of **languages and dialects of Italy** through responsible, participatory, and locally-meaningful development of language and speech technology*

- Introduce newcomers to the *speaker-centric* approach
- Foster *discussion on practices* in diverse environments
- Encourage *participatory work* between fields of study, speech communities, and cultural institutes
- Raise *awareness* on the linguistic heritage of Italy

Illustration by rawpixel.com / freepik

Alternative directions

Regional Italian in NLP

- Account for it *at all levels* (not lexical only!), investigate *disparities* across variants

Linguistic atlases and dialectometry

- Collaborate w/ *atlas projects*, complement *qualitative* studies with *quantitative* work

Functional, social, and contextual aspects of code-switching

- Study the *why* and *when* towards understanding language replacement processes

Conclusion

- Language varieties of Italy have diverse ***functions*** and ***contexts***
- Speech communities have different ***needs*** for technology
- ***Variation*** is *natural*, and *language* and ***culture*** are inseparable
- ***Engaging*** with communities: great opportunities for work that *matters*