Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-025-05025-1 OPEN

Extracting toponyms from OpenStreetMap and other gazetteers: comparing representational accuracy in multilingual contexts

Francesco-Alessio Ursini[1]✉ & Giuseppe Samo[2]✉

The goal of this paper is to investigate OpenStreetMap as a research tool by analysing what *pros* and *cons* this platform offers to linguistics and to GIS disciplines. To reach this goal, the paper analyses how this platform represents places as geographical units and toponyms (i.e. place names) as linguistic units referring to places. The paper presents two previous studies that featured a novel procedure for toponym extraction and its application to OpenStreetMap toponym data. These two studies focused on distinct scales and densities of geographical distribution in multi-lingual contexts: city level (Macao); mixed regional and national level (Italy). The studies also included a comparison of these data with data originating from an authoritative geographic source (e.g. Italian street directories). The present paper extends the analysis and results from these studies by showing that via a single extraction algorithm, one can obtain all the relevant toponyms from overpass-turbo, a platform including OpenStreetMap's textual information, and from other gazetteers. For each level of analysis, the paper shows that toponyms come in different combinations of multi-lingual formats: Chinese and Portuguese for Macao, Italian, local dialects (e.g. Genoese), and minority languages (e.g. German) for Italy. From these data, the paper offers an analysis of language-specific features, methodological challenges, and informational accuracy of each database. The paper proposes that OpenStreetMap may be as reliable as authoritative sources; however, one must apply cross-source comparison during data analysis, to confirm OpenStreetMap-based data. The paper concludes by discussing the current role of OpenStreetMap as an information database in toponym extraction. The paper discusses the use of OSM in linguistics and GIS disciplines, and how these uses can offer theoretical insights informing research in these disciplines.

[1] Central China Normal University, Wuhan, China. [2] Beijing Language and Culture University, Beijing, China. ✉email: randorama@outlook.com; samo@blcu.edu.cn

Introduction

Geographic Information Science (henceforth: GIS) and Linguistics may initially appear as unrelated disciplines. In highly schematic terms, GIS investigates various types of phenomena from the perspective of their geographical and spatial distribution (Fotheringham and Wilson, 2007). Linguistics, instead, investigates languages and their distinctive properties (Fasold and Connor-Linton, 2014). However, linguistics includes sub-disciplines such as geo-linguistics, dialectology, and toponomastics, which respectively study the geographical distribution of languages, dialects, and place names or *toponyms* (e.g. Perono Cacciafoco and Cavallaro, 2023). Conversely, several sub-disciplines in Geography and GIS study toponyms, their use and socio-cultural status across different cultures and languages (e.g. Alderman, 2022; Gnatiuk and Melnychuk, 2020). Thus, GIS and linguistics overlap in their domains of enquiry and research methodologies when focusing on spatial data, broadly defined. They then seem to converge in their focus on toponyms as a source of data regarding human understanding of spatial information, also broadly defined.

An open question is whether these disciplines can also share information sources from which they can extract their toponymic data. Recent linguistically oriented works have offered a preliminary positive answer by using OpenStreetMap (henceforth: OSM[1]) to extract a large set of toponyms and analyse their grammatical properties (Ursini and Samo 2023). However, the work does not offer a detailed analysis of how information about toponyms and their properties appears in OSM. Therefore, this and other similar works do not address the theoretical and methodological problems that emerge once researchers extract, process, and manage data from this source, and compare these data with data from official sources (e.g. gazetteers and land registry data).

The goal of this this paper is to analyse OSM as a research tool and data source for linguistics and GIS, thus comparing this source with official sources. In so doing, we also aim to show that OSM can be a useful inter-disciplinary source for linguistic and geographic research on toponyms (e.g. respectively, Perono Cacciafoco and Cavallaro, 2023; Rose-Redwood, Alderman and Azaryahu, 2018). We present two case studies in which we carried out toponym extraction and analysis in multi-lingual contexts defined at different scales and densities of geographic distribution (city level, Macao; regional and national level, Italy). We analyse the methodological problems emerging from this extraction procedure, the type of linguistic data and the degree of empirical coverage of OSM when compared with other sources. We thus aim to show that several disciplines studying toponyms can amply benefit from using the OSM source, in combination with other accessible sources.

We organise our paper as follows, to reach this goal. We first offer an overview of OSM and previous OSM-based works on toponyms, thus introducing three research questions (Section 'Literature review: previous research on OSM and current challenges'). We then present our methodology and materials (Section 'Methodology and materials'), and then the specific studies and results by which we answer each research question (Section 'Results'). Section 'Discussion' offers a discussion as a general answer to our research questions; Section 'Conclusions' concludes.

Literature review: previous research on OSM and current challenges

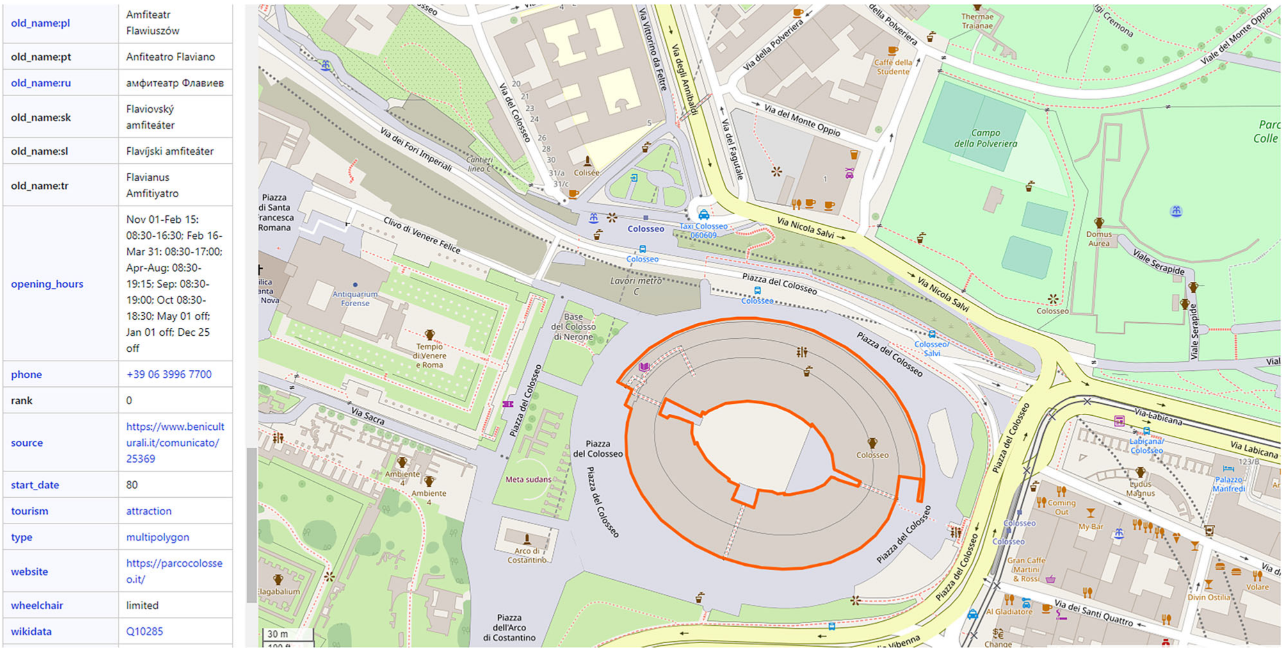
OSM is an online platform that offers 'a free, editable map of the world' (Curran et al., 2012, 2013; Keßler, 2017). Since its online appearance in 2004, OSM has provided open-source, easily editable maps of increasing detail and definition to all users and contributors (Arsanjani et al., 2015a; Mooney et al., 2017). OSM founders based the platform on a philosophy known as volunteered geographic information (henceforth: VGI, Antoniou and Skopeliti, 2017; Goodchild, 2007; Keßler et al., 2009; Sui and Goodchild, 2011). Any registered user can become a contributor by inserting and editing information regarding locations and the objects occupying these locations. OSM has thus emerged as an important source of knowledge for researchers in GIS and other disciplines focusing on geo-spatial information (e.g. urban planning, data mining), due to its flexibility and ease of management.

Contributors can enrich OSM maps with spatial information; however, its core geographical objects work as follows (Almendros-Jiménez et al., 2021; Rajšp et al., 2021). Registered contributors can edit information based on their knowledge of locations. Contributions centre on the geographical objects shaping maps: *nodes* representing locations, ways representing connections among locations, and relations between nodes and/or ways. Each object has tags, labels indexing attributes ('keys') and values associated to locations (e.g. coordinates, altitude, shape, type of location). Tags represent objects on maps via a dual visual and textual format. Visually, tags are icons for objects on maps; textually, tags are key-value matrices of the type found in linguistics, GIS, and other computational disciplines (Gamerschlag et al., 2015; Sag, 2012). Tags are therefore unique multi-modal (i.e. visual and textual) indexes, as illustrated in Fig. 1 (left panel):

Contributors can introduce icons, keys and values according to their knowledge of a location and the objects that occupy this location; guidelines and graphical tools can streamline this process. For instance, local contributors from a neighbourhood can insert information pertaining to two buildings that are still unreported in OSM. They can create two new objects and tags, fill the tags with sets of keys describing the buildings, and select 'building' icons (i.e. visual tags; Salvucci and Salvati, 2022; Zhou et al., 2022). Contributors can also insert keys and values for ways (e.g. streets connecting to these buildings), and the informational content of relations. For instance, buildings can operate as habitations for citizens; streets may be quite busy during rush hour, and so on. Contributors can continually update tags that represent the physical-geographical properties of objects, but also the possible relations between these objects and the individuals interacting with the objects (Mayer et al., 2022).

OSM tags can therefore offer information about *places*: geographical objects in which humans perform activities and to which they can possibly develop forms of social, cognitive, and psychological attachment (Cresswell, 2014; Malpas, 2018; Tuan, 1977). Toponyms can consequently act as names that carry this complex, partially subjective information via their semantic content and ability to refer to places and their attachment relations to human individuals (Blair and Tent, 2015, 2021; Perono Cacciafoco and Cavallaro, 2023). Nodes and ways can be objects representing places of increasing complexity: from buildings to cities and regions, and from streets to highway networks. Irrespective of this complexity, they can represent places and the rich informational content that contributors associate with places. OSM can thus operate as a multi-modal map integrating both spatial and 'platial', i.e. place-based information (Arsanjani et al., 2015; Mayer et al., 2022).

The focus on platial information has allowed GIS researchers to use OSM as a data source for several topics. OSM maps can include places as small as ATMs, trees, and benches (Touya et al., 2017). Maps can also provide real-time information regarding risks affecting places (e.g. natural disasters: Cerri et al., 2021; Hecht et al., 2013; Seto, 2022); epidemic diffusion (Mooney et al., 2021; Mooney and Juhász, 2020). OSM maps can then provide online information about real-time updates that contributors



old_name:pl	Amfiteatr Flawiuszów
old_name:pt	Anfiteatro Flaviano
old_name:ru	амфитеатр Флавиев
old_name:sk	Flaviovský amfiteáter
old_name:sl	Flavijski amfiteáter
old_name:tr	Flavianus Amfitiyatro
opening_hours	Nov 01-Feb 15: 08:30-16:30; Feb 16-Mar 31: 08:30-17:00; Apr-Aug: 08:30-19:15; Sep: 08:30-19:00; Oct 08:30-18:30; May 01 off; Jan 01 off; Dec 25 off
phone	+39 06 3996 7700
rank	0
source	https://www.beniculturali.it/comunicato/25369
start_date	80
tourism	attraction
type	multipolygon
website	https://parcocolosseo.it/
wheelchair	limited
wikidata	Q10285

Fig. 1 Tags for the Colosseum in Rome. The visual tag (i.e. the Colosseum's location) is the grey oval shape in the centre of the figure.

perform on objects (e.g. fires in buildings: Novack et al., 2022; Schäfer and Kieslinger, 2016; Senaratne et al., 2017). Recent works have thus suggested that OSM maps are becoming increasingly spatially, temporally, and platially accurate. They therefore offer a dynamic view of the places they represent (Mocnik, 2022; Romm and McKenzie, 2023), to the benefit of casual users and geo-spatial scholars alike.

The fact that OSM can provide dynamic information updates and management tools has, however, raised certain research questions. Traditionally, research in GIS has centred on gazetteers, ordnance survey maps and other representational sources produced by official geographic institutions, i.e. authoritative geographic information sources (henceforth: AGI Bortolini and Camboim, 2019; Fize et al., 2021). Geographic institutions mostly adopt top-down practices of information collection and management. For instance, officials from the land registry can register information about buildings and streets from local authorities owning these places. Local citizens may never be involved in these procedures, even if they have knowledge pertaining to the aforementioned places. OSM stakeholders, instead, have mostly adopted bottom-up management practices. Contributors are usually citizens interested in inserting information about their local places, often addressing coverage gaps cropping up in AGI sources (Keßler, 2017; Keßler et al., 2009).

Overall, contributors' knowledge may be accurate to sometimes-volatile degrees, thus casting a shadow on the accuracy of VGI sources. Several studies have, however, shown that the soundness of spatial information in VGI sources strongly correlates with contributors' formal education, motivation, and commitment to professional-like data insertion (Garba et al., 2022; Holthaus and Thiemermann, 2022; Jaljolie et al., 2023). Furthermore, several AGI sources have become freely accessible to the public, and thus accessible to OSM contributors for systematic 'information dumps', i.e. massive imports from other sources (Bravo and Sluter, 2022; Wu et al., 2022). Corporations and NGOs have also begun to hire professional contributors performing large data sets imports, since business and to citizens' communities consider the support of OSM beneficial (Anderson and Sarkar, 2020; Sarkar and Anderson, 2022). Hence, OSM is becoming a 'multi-source' model of information management, in which contributors reconcile top-down and bottom-up philosophies via carefully documented data (Hu et al., 2022).

The importance of this multi-source model in platial analysis becomes clear when one focuses on toponyms. In highly schematic terms, OSM tags usually include a 'name' key (i.e. attribute), among their many keys. The specific value for this key is usually the toponym assigned to a given place. In our example involving the new buildings, we can have an OSM tag for each building, including near-identical values, in case the same company built both buildings. However, we can have one building called 'Joseph Joestar', and the other 'Jotaro Kujo': toponyms are usually unique, distinctive labels for places. Citizens interacting with these buildings and OSM will likely exploit this platial uniqueness to refer to either building. For humans, toponyms are more cognitively accessible than pure spatial information (e.g. coordinates, lists of features: Perdana and Ostermann, 2018). This is the case because they are the key language category that allows humans to talk about places (Alderman, 2022; Perono Cacciafoco and Cavallaro, 2023; Rose-Redwood et al., 2018).

One can thus define the central role of toponyms in OSM as follows (cf. again. Mocnik, 2022; among others). Toponyms act as prime keys leading to the access of the complex semantic information defining place descriptions. Any contributor can insert an official or unofficial toponym for a place, and provide references/sources from which this information originates (e.g. personal knowledge, other volunteer-based sources, and authoritative sources). Other contributors can question the reliability of this information via their own sources, and one can solve eventual disputes via cross-referencing and online discussions aimed at avoiding 'editing wars'. Information about toponyms can receive updates in real time, as in the case of information about places. Toponyms in OSM are therefore platial data types that are as important as geographic data types, due to their immediate cognitive appeal to general users and researchers alike.

In recent times, several studies have used OSM for toponym analysis, exploiting its multi-source model of data integration (Ahmadian and Pahlavani, 2022; Hall and Jones, 2022; Kaisar Ahmed, 2022; Machado et al., 2021). For instance, contributors to the Paris' toponym database have integrated grassroots knowledge with information from public gazetteers (Antoniou et al., 2016). The Jerusalem database includes coverage of Hebrew toponyms hinging on gazetteers' imports, but coverage of Palestinian toponyms has gained momentum, even if Palestinian users cannot rely on AGI sources (Carraro, 2021). Notably, OSM contributors tend to focus on Europe and North America. However, coverage of other countries is increasing at a dramatic, if uneven pace (e.g. Brazil: Kaisar Ahmed, 2022; China: Qian et al., 2016; Kenya: Daniel and Mátyás, 2022). Contributors generally work intensely on the task of assigning names to each object perceived as a place.

OSM is thus becoming a reliable even if still partially unbalanced, resource for GIS studies. Its role in linguistic research, however, appears to involve two apparently distinct problems. The first problem can be pre-theoretically defined as a problem of heterogeneous distribution. Recent publications show that coverage of platial information tends to be denser at smaller, local scales (Westerholt, 2019a, 2019b). Contributors to VGI sources may offer coverage of the districts or cities they live in. However, the distribution of platial information at a regional and national level may include vast regions of missing information. For instance, rural zones and underdeveloped urban zones tend to correlate with poor toponym coverage (Daniel and Mátyás, 2022; Elias et al., 2023; Qian et al., 2016). Toponyms' spatio-temporal density in OSM can therefore reflect a region's salience or irrelevance for societies and populations. A consequent question is how OSM compares to AGI sources, with respect to toponyms' distribution.

The second problem can receive a formulation as a problem of *heterogeneous multi-lingual representation*. Toponomastics and critical toponymy have shown that different communities may decide to name places via different socio-cultural practices (e.g. Cavallaro et al., 2019; Stolz and Warnke, 2018). These practices may, however, follow conflicting guidelines and involve subtle power conflicts, in multilingual contexts (Alderman, 2022; Azaryahu, 2011; Gnatiuk and Melnychuk, 2020; Rose-Redwood et al., 2010). For instance, *Uluru* is the sacred toponym that local Australian Aboriginal communities use for the monolith in the centre of Australia. For most people, however, the English Australian name *Ayers Rock* may be more familiar. Both are eponymous names: however, the former is based on a divinity's name while the latter is based on the name of a previous governor of the South Australia state. Crucially, the Jerusalem case suggests that co-existing linguistic communities can have uneven access to OSM, due to language-external pressures on accessibility (Carraro, 2021). A consequent question is how AGI sources and OSM differ in multilingual coverage, due to these accessibility asymmetries stemming from socio-linguistic and geo-linguistic factors.

As matters stand, these two intertwined problems of heterogeneity lead to the emergence of three compound research questions. The first research question arises from the distribution problem; the second question, from the multi-lingual problem; the third question, from their theoretical implications. The three research questions can receive the following formulations:

- **RQ1**: How many toponyms can one find in OSM, and how accurate and homogeneous this coverage is? How does OSM compare with authoritative sources?
- **RQ2**: What asymmetries can one find in OSM, when one analyses the coverage of toponyms across the multiple languages used in a delimited region? Where these asymmetries emerge, and at what scales of analysis they emerge? How OSM compares with AGI sources?
- **RQ3**: How results based on OSM can inform GIS research, toponomastics, and other sciences studying toponyms and their properties?

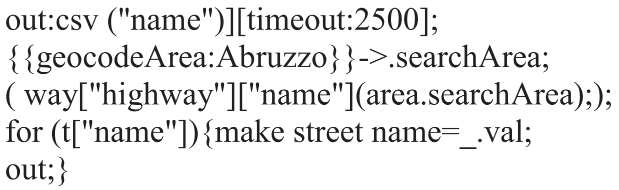
We answer **RQ1** from a mostly quantitative perspective. We thus analyse the distribution of toponyms at three levels of geographical scale and density: the city, regional and national levels. We propose two case studies: Macao, in China (city level), plus Italy and its 20 regions (national, regional levels). We then discuss how these databases include toponym information from AGI sources (e.g. official gazetteers). We answer **RQ2** from mostly a qualitative perspective. We thus analyse the differences in multi-lingual coverage across the official languages of each target study (e.g. Chinese and Portuguese in Macao), and the toponyms attested in these regions. We then discuss where and at what scales these asymmetries arise. We answer **RQ3** by integrating these two perspectives into one model, and by discussing how one can compare OSM data to AGI sources' data. We subsequently discuss how OSM data can find applications in linguistics, GIS, toponomastics, and other sciences focusing on platial information.

Methodology and materials

We used one language-general methodological approach to data extraction and processing; however, we discuss language-specific adjustments in section 'Results'. Portions of each study appeared in previous works that analysed toponyms from a toponomastic perspective (Xie et al. 2023, Samo and Ursini 2023). In this study, we present a broader range of geographic and geo-linguistic data to address the first two research questions and a more in-depth meta-analysis of the data to address the third question (cf. Ursini and Samo 2023). We acknowledge that a focus on one geographical domain (e.g. Italy) could have been a more practical and empirically coherent choice. A methodological goal motivated this less practical choice, however. By using our two previous studies as a baseline, we can indirectly confirm that other researchers can replicate our methodology irrespective of the geographical region and scale under discussion. We thus trade higher data cohesiveness with evidence about the repeatability of the procedure.

We accessed OSM data through the platform overpass-turbo[2], sizing our search in the relevant geographical areas. We queried the platform with the script in Fig. 2 to extract the relevant data (in the example, to extract toponyms in the 'Abruzzo' Italian region). We proceeded with data analysis following the flowchart in Fig. 3:

As the flowchart shows, the data extraction step generates a raw data file that we transformed in a.csv format ('OSM Turbopass' step). The output.csv file (i.e. the materials) easily supports statistical data analysis, visual data representation, and linguistic categorisation. The additional step in this paper with respect to the two previous works is that we focus solely on generic terms in



```
out:csv ("name")][timeout:2500];
{{geocodeArea:Abruzzo}}->.searchArea;
( way["highway"]["name"](area.searchArea););
for (t["name"]){make street name=_.val;
out;}
```

Fig. 2 An example of the algorithm used for data extraction. Our query retrieves data from OSMP within the specified geocode area (in the first line of our example, the Italian region of 'Abruzzo') in each timeout time specified in milliseconds. It then looks for toponyms via the 'highway' label, and for their tags (second line). The query extracts the toponyms (third line), and outputs the results in alphabetical order in CSV format (fourth line). The extraction procedure covers not only ways/highway toponyms, as the label seems to suggest, but also other toponym types within a given area. See the main text for discussion on the types of toponyms extracted in the studies.

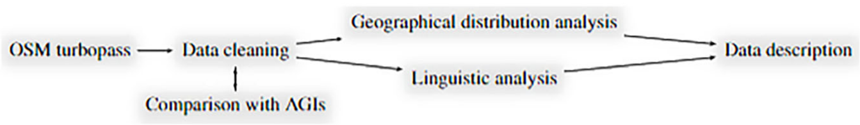
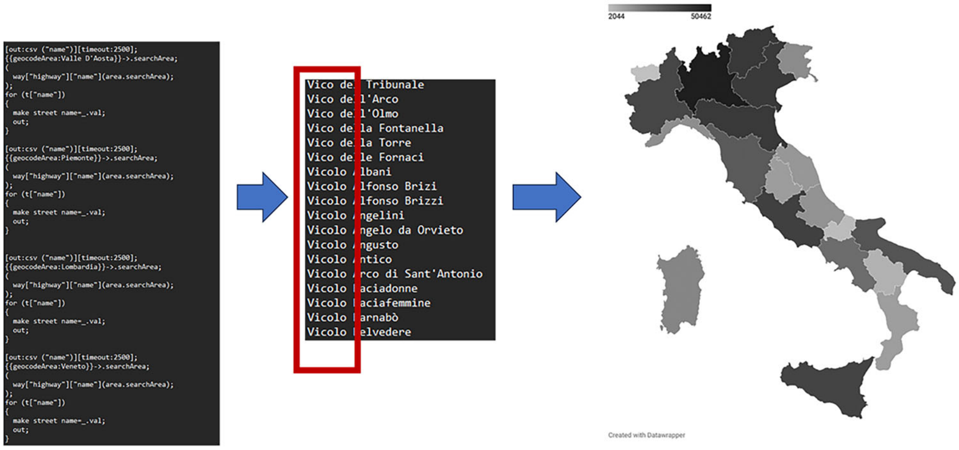


Fig. 3 Flowchart representing the methodology used in each study.



```
[out:csv ("name")][timeout:2500];
{{geocodeArea:Valle D'Aosta}}->.searchArea;
(
  way["highway"]["name"](area.searchArea);
);
for (t["name"])
{
  make street name=_.val;
  out;
}

[out:csv ("name")][timeout:2500];
{{geocodeArea:Piemonte}}->.searchArea;
(
  way["highway"]["name"](area.searchArea);
);
for (t["name"])
{
  make street name=_.val;
  out;
}

[out:csv ("name")][timeout:2500];
{{geocodeArea:Lombardia}}->.searchArea;
(
  way["highway"]["name"](area.searchArea);
);
for (t["name"])
{
  make street name=_.val;
  out;
}

[out:csv ("name")][timeout:2500];
{{geocodeArea:Veneto}}->.searchArea;
(
  way["highway"]["name"](area.searchArea);
);
for (t["name"])
{
  make street name=_.val;
  out;
}
```

```
Vico del Tribunale
Vico dell'Arco
Vico dell'Olmo
Vico della Fontanella
Vico della Torre
Vico delle Fornaci
Vicolo Albani
Vicolo Alfonso Brizi
Vicolo Alfonso Brizzi
Vicolo Angelini
Vicolo Angelo da Orvieto
Vicolo Angusto
Vicolo Antico
Vicolo Arco di Sant'Antonio
Vicolo Baciadonne
Vicolo Baciafemmine
Vicolo Barnabò
Vicolo Belvedere
```

Fig. 4 An example of the process for toponyms' extraction from the second study, which investigates the distribution of local toponyms in Italy for given regions. The script's output is a csv file and the data become plotted in maps (the rightmost panel is the distribution of number of tokens of toponyms across the regions of Italy, map created with Datawrapper v.1.25.0 Lorenz et al., (2012). Please consult the lists of script/queries used for data extraction in the supplementary files, 'List of Queries A' and 'List of Queries B' files.

toponyms. We define generic terms as the terms that describe/classify a place carrying a given name (e.g. the term *vicolo* 'alley' in *King's alley*: Blair and Tent, 2015, 2021). This type of analysis provides crucial details on the geographical distribution of items and thus yields quantitative evidence, as we show in Section 'Methodology & Materials'. Figure 4 illustrates how we first converted the raw data files to.csv files, how we extracted generic terms, and then how we prepared the visual maps for the data:

Once the.csv file was ready ('Data cleaning' step), we compared these data with the data from selected AGI sources for each study ('Comparison with AGIs' step). We then plotted the distribution and density of data ('Geographical distribution analysis' step), and performed an analysis of the grammatical and typographic properties of toponym sets in each study ('Linguistic analysis' step). From these different components of the study, we created a description of the data ('Data description' step), which forms the section 'Results'. We therefore use mostly geographic data to answer **RQ1**, linguistic data to answer **RQ2**, and their combined model to answer RQ3. We clarify further intermediate steps in the linguistic and geographic distribution analyses once we present each case study.

Before we move to the description(s) of the data, we offer two methodological clarifications and one clarification about the materials. First, the comparison of the OSM data with AGI data achieves a form of what psychologists define as triangulation, i.e. the analysis of the same dataset(s) via multiple methods and/or sources (Damico and Tetnowski, 2014; Rothbauer, 2008). In GIS, there is also a growing awareness that toponym retrieval and analysis studies must involve multi-source methods implementing forms of cross-verification (Hu, Al-Olimat, et al., 2022; Hu et al., 2022; We thus assume that by implementing a form of methodological triangulation, we increase the reliability of our findings, and properly compare OSM as a VGI source with AGI sources. Second, we focus on toponyms as one data type. When relevant, however, we explain how toponyms map onto the other data types forming tags (e.g. spatial coordinates).

The clarification about the materials pertains to the sub-type of toponyms we extracted in each study. We extracted toponyms describing places in urban administrative zones, known in the literature as 'urbanonyms' (Vannieuwenhuyze, 2007; David, 2011; Way, 2019; Xie et al. 2023, Samo and Ursini 2023). Hundreds of works have studied toponyms for streets (hodonyms), toponyms for squares (agoranyms), and other sub-types of urban toponyms, especially in critical toponymy (e.g. Alderman, 2022; Azaryahu, 2011; Rose-Redwood et al., 2010, 2018; Gnatiuk and Melnychuk 2020). We cannot possibly review all the relevant literature here, but one can find recent state-of-the-art overviews are in Coates (2007); Basik (2020, 2021); Walkowiak (2024). In this paper, the tokens forming the analysis are toponyms for streets, squares, parks, points of interest (e.g. monuments), and other places forming the urban zones under consideration. We further clarify relevant study-specific details along with the results. We then address the methodological theoretical import of these results in the discussion section.

Results

We present the results of each study and our study-specific answers to **RQ1** and **RQ2** in this section. The data for the studies are in the supplementary materials (Supplementary file A, B for the first Study; Supplementary file C for the second study).

First study: Macao, China (locally bilingual, city level). In the first study, we analysed the urban toponyms from Macao, a city, and special administrative region (SAR) in South-East China. Macao has a centuries-long tradition of multi-lingualism, as a former Portuguese colony. European settlers introduced Portuguese has the official language of their rule; this language co-existed with Cantonese, the Sinitic language spoken in the Guangdong province (Yee, 2014). In modern Macao, Portuguese has remained as an official language, though only 1% of the population speaks it natively. Cantonese and other Sinitic languages (e.g. Mandarin and Hakka) are prevalent and English is slowly becoming a *de facto lingua franca* in Macao, for economic purposes (Botha and Moody, 2021). The first study thus provided a complex multi-lingual environment at a city level/scale.

Gazetteers include toponyms in Chinese and Portuguese, the two official languages. Chinese toponyms are written in Chinese simplified characters, and are thus intelligible to speakers of any Sinitic language (e.g. Chinese/Mandarin, Cantonese and Hakka). Portuguese toponyms are written in Macanese Portuguese, which is nearly identical to standard European Portuguese. The authors analysed the grammatical and lexical properties of each token toponym and corresponding general term. One author was a native speaker and reader of Mandarin, and another author had a high degree of fluency in Portuguese. Chinese generic terms provided a minor challenge when they appeared as via two-character compounds (e.g. 公園 *gung1 jyun4* ‘public garden’). The paper’s native author solved this challenge by assessing whether such compounds would jointly describe a distinct place type. See Supplementary file A for the data regarding the results from the linguistic analysis performed in Xie et al. (2023).

We compared the resulting sets against an official gazetteer from the Macao government in CD-ROM form, as our AGI source (Cartography and Cadastre Bureau of Macau SAR, 2021). This gazetteer includes names for streets, squares, parks, POI’s, and other places within the urban administrative territory of Macao. Thus, these toponyms can qualify as urban toponyms irrespective of the type of place they name (Xie et al. 2023). The gazetteer represents an AGI source because the Macao government handles administrative matters regarding Macanese toponyms, and updates online and off-line maps (e.g. CD-ROM releases approximately at bi-annual intervals). For the current study, we collected further data with respect to (Xie et al. 2023) to address geo-distributional and geo-linguistic aspects. The three key results are as follows.

First, we obtained two lists of 1394 toponyms from both sources (OSM, CD-ROM). We compared the two sources by using the Jaccard Index of similarity (from 0 to 1: the closer to 1, the more similar two populations, Jaccard, 1901), and obtained 0.989 as a result. OSM toponyms only presented minor spelling variants in Portuguese that stem from contributors’ mistakes (e.g. accent omission: *Rua de Santo Antonio* instead of *Rua de Santo António* 'Saint Anthony Street’). Notably, the etymology of toponyms often differed from Portuguese to Chinese, due to different naming practices in each linguistic community (45% of the total). For instance, 龍鬚街*lung4 sou1 gaai1* ‘Dragon Beard Street’ and *Rua Central* ‘Central Street’ are the two toponyms for a key street in Macao’s old town. Portuguese authorities named this street after is location and social function; Chinese speakers, after the imagined appearance of a local temple.

Second, we compared the two data sets on a place-by-place basis. We thus analysed the geographic distribution of these toponyms on the Macao territory, while also analysing the density of urban constructions and agglomerates across Macanese districts. Our conjecture was that zones with higher numbers of human-built places (e.g. buildings, streets and squares) would also feature a higher number of toponyms (cf. Hecht et al., 2013; Salvucci and Salvati, 2022). We then verified that each Portuguese-Chinese toponym pair uniquely corresponded to one place via an analysis of the toponyms’ coordinates. We present the geographical distribution of these toponyms in Fig. 5 (extracted from Xie et al. 2023):

As the map indirectly shows, the geographical distribution of toponyms correlates with the degree of urban development. For instance, the old town has the highest density of toponyms because it includes the highest number of places with key social functions in Macanese society (e.g. the Senate building). Instead, the southern island of Coloane ('Ilha de Coloane', in the map) is still mostly composed of historical spots (e.g. temples), natural reserves and scenic views. One can find toponyms along its coasts, where these spots and views are located. Therefore, Macanese toponyms may have a heterogeneous distribution as a reflection of places’ size and spatial prominence (cf. Elias et al., 2023; Kaisar Ahmed, 2022).

Third, we compared the OSM data with other AGI sources, and analysed whether OSM updates had occurred from the time of the original study. We first consulted the APP place directory for Macao provided by the ministry for tourism.[3] The APP implements data from official gazetteers for Portuguese and Chinese. It also includes English toponyms as direct calques (i.e. translations) from Portuguese. The 1633 toponyms thus also include toponyms for casinos, historical buildings and other 'points of interest' (POI’s). For instance, English *A- Má Temple* is a translation of Portuguese *Templo de A-Má*, and Chinese 媽閣廟*maa5 gok3 miu6*. English features in this APP because this language is slowly emerging in official toponymy (e.g. in street plaques in the old town: Botha and Moody, 2021). The APP thus seems to reflect this increasing relevance of this third language.

Table 1 offers an overview of the extra English toponyms one can find in the APP by listing the generic terms only found for the toponyms exclusive to this APP:

The map in Fig. 6 shows that OSM includes almost all the casino toponyms attested in the APP: four toponyms appear missing (cf. OSM’s 35 units vs. the APP’s 39). Their respective toponyms mostly occur in the old town and in Taipa, the northern and more urbanised side of the southern island; no casinos exist in Coloane. The data presented via Figs. 5, 6 therefore suggest that places and hence toponyms tend to occur in the densely urbanised parts of Macao, i.e. the city’s zones with a more intense human presence. As we now have a broad set of data at disposal, we can answer our first two research questions with respect to the city level of platial distribution.

Regarding **RQ1**, OSM now includes a slightly higher quantity of data than the official CD-ROM gazetteer. However, the quality of this data is marginally inferior for the Portuguese dataset. An analysis of the updates’ history suggests that one update occurred from the date of the original study. This update brought OSM’s accuracy above the CD-ROM’s level, but below the tourist office APP’s level. Furthermore, the density of toponyms on the Macao maps from all three sources (OSM, official gazetteer, tourist office APP) supports the correlation between human presence and the heterogeneous distribution of places. Zones with dense human presence attract dense, potentially homogeneous clusters of toponyms; zones with scarce human presence attract rarefied, potentially heterogeneous clusters. Thus, OSM appears to offer a slightly more homogeneous and accurate coverage of toponyms than one AGI source (the CD-ROM gazetteer), but a less accurate coverage than another AGI source (the tourist office APP).

Regarding **RQ2**, OSM reveals multi-linguistic information about toponyms in Portuguese and Chinese. Crucially, this multi-lingual coverage is as accurate as the official CD-ROM gazetteer. The tourist office APP also includes English toponyms, since it works as a gateway language for tourists visiting this city, and aims to capture the growing relevance of this language in Macanese society. Thus, the APP acts as an AGI source that presents a broader, multi-lingual and detailed mapping of Macanese toponyms than OSM. Nevertheless, one can assume that OSM contributors are already implementing information from this source in ongoing updates (cf. the casino data). These data overall confirm that OSM can provide relevant evidence for linguistic analyses due to its multi-lingual status. However, OSM can involve missing data involving specific toponym types (e.g.

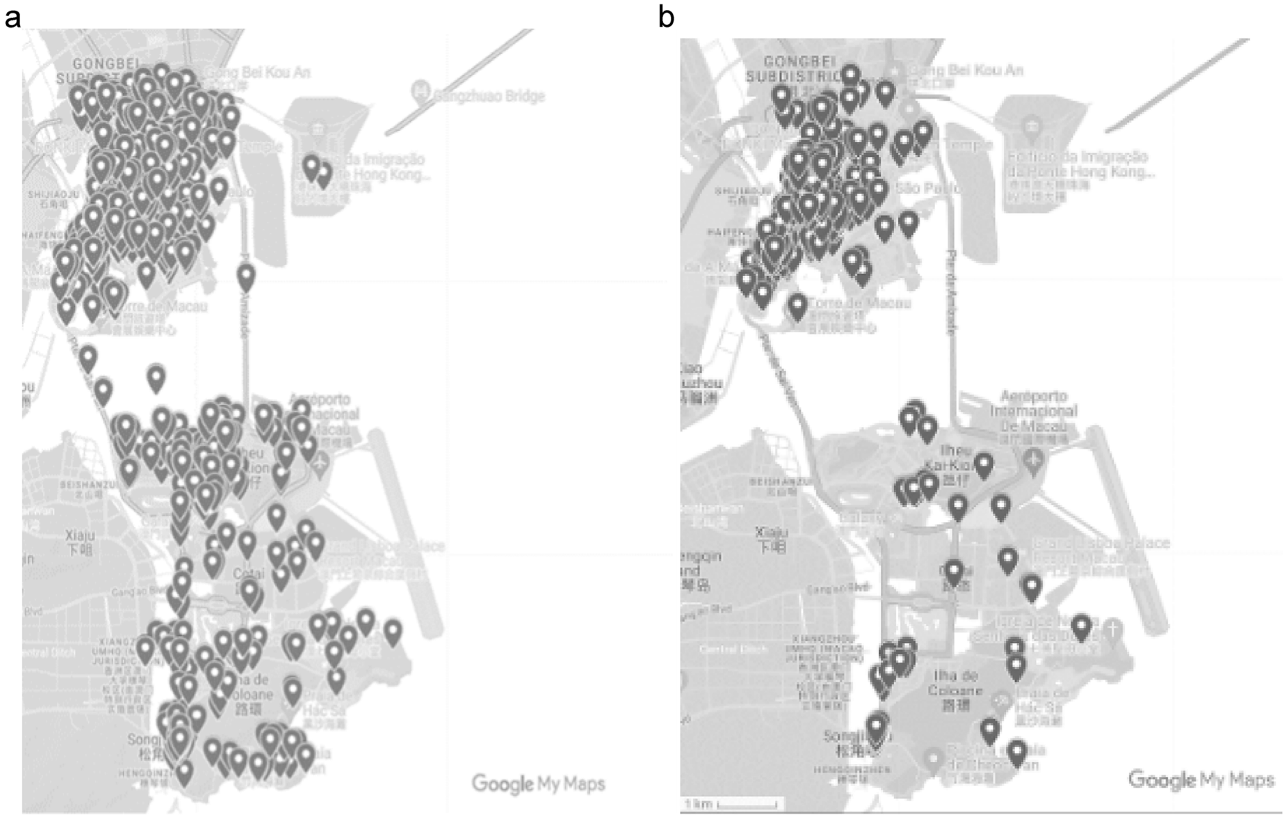


Fig. 5 Distribution of near-equivalent terms. a We present near-equivalent terms (see details in Xie et al. 2023), namely 'near-equivalent' 1-to-1 translations of the generic terms in Chinese and Portuguese. As a control group, **b** The distribution of dissimilar terms representing instances in which the Portuguese and the Chinese terms are not near-equivalent translations. Extracted from Xie et al. (2023, p. 37) Fig. 1. See also the 'List of Macau Streets' file in the supplementary file section for the full list of urban toponyms used in this analysis.

POI toponyms) at local scales of analysis. AGI sources may still provide more homogeneous and accurate coverage of toponyms.

Second study: Italy and its regions (multi-lingual, regional and national level). In the second study, we investigated the potential dialectal origins of Italian urban toponyms via OSM (Samo and Ursini 2023). In Italy, standard Italian co-exists with geographical dialects that are considerably different from this language. Differences involve socio-linguistic aspects (e.g. Berruto, 2012), grammatical, phonological and lexical features (e.g. Bossong, 2016; Samo and Ursini 2023). For this study, two basic considerations play a role. First, linguistic research considers dialects such as Neapolitan in the South, Piedmontese in the North distinct languages, since their linguistic features are different enough from Italian to warrant this status. Italian and most languages/dialects, however, belong to the Romance branch of Indo-European languages. They have interacted over the centuries, as closely related languages. Second, toponyms may nevertheless originate from languages other than Italian (e.g. German, French: Cassi and Marcaccini, 1998; Marcato, 2009). The second study and the results provided in this study thus provide evidence for a situation of nuanced multi-lingualism.

We offer a concise overview. Italian toponyms often originate in the languages of the pre- Italic populations that once inhabited Italy (e.g. Etruscan), but also in the local dialects (e.g. Florentinian in Florence; Chiappinelli, 2013; Cassi, 2015). The legislation for Italian 'odonimi' (i.e. urban toponyms, Mastrelli, 2005) establishes that local administrations (e.g. municipalities) assign toponyms to urban places (e.g. streets), also via the consultation of local citizens. In regions with special administrative status (Valle D'Aosta, Trentino-Alto Adige), parts of the population have an official status as 'linguistic minorities'. These communities form less than half of the population and speak a different language from the official language (Mastrelli, 2005). In these regions, toponyms occur in both official languages (Italian and French, Italian and German), with the minority language preceding Italian (e.g. Bozen/Bolzano, German/Italian, for the administrative seat of Alto Adige).

Building on these premises, Samo and Ursini (2023) used OSM to analyse urban toponyms and their generic terms. The study showed that toponyms including generic terms not attested in standard dictionaries of Italian or glossaries of geographic terms (e.g. Calafiore, 1975; Gasca Queirazza et al., 1990; De Mauro, 2020) may be dialectal in origin. Such urban toponyms often entered the modern Italian language via a process of spelling standardisation and lexical absorption. For instance, in the city of Genoa, one can find terms such as *crosa* in toponyms for the crimson alleys traversing the city's quarters. However, the Genoese term was originally *creusa* (e.g. Italian *Crosa del Mare* from Genoese *Creusa de Ma'*). Crucially, Genoese ('Zeneize', in the original language) is a Gallo-Romance (i.e. Francophone) dialect/language mostly spoken in the Liguria region (Bossong, 2016). The study thus showed that Italian includes hundreds of generic terms, and the toponyms they occur in, from local geographical dialects/languages.

The second study concentrated on toponyms for places being part of urban administrative zones (Way, 2019). Thus, its tokens include toponyms, for e.g. streets, squares and range from toponyms belonging to urbanised villages (e.g. Chiusi, in Tuscany) to toponyms from the capital city, Rome. For the present study, we extracted the set of generic terms attested in urban toponyms in

Table 1 English generic terms, and their Portuguese and Chinese Counterparts (adapted from Xie et al. 2023, page 38, Table 4).

English (tokens)	Portuguese	Chinese
Leisure Area (75)	*Zona de Lazer*	休憩區*jau1 hei3 keoi1* 'rest area'
Casino (39)	*Casino*	酒店*zau2 dim3* 'hotel'
Temple (30)	*Templo*	閣 *gok3* (廟 *miu6*) 'pavilion (temple)'
Museum (26)	*Museu*	博物館*bok3 mat6 gun2* 'museum'
Library (21)	*Biblioteca*	圖書館 *tou4 syu1gun2* 'library'
Church (9)	*Igreja*	堂*tong4* 'hall'
Square (9)	*Largo*	前地 *cin4 dei6* 'front place'
Fortress (8)	*Fortaleza*	炮 台 *paau3* toi4 'fortress', 教堂 *gaau3 tong4* 'church' 燈 塔 *dang1taap3* 'lighthouse'
Building (3)	*Edifício*	局*guk6* 'bureau'
Park (3)	*Parque*	公 園 *gung1jyun4* 'public garden'
Corridor (2)	*Acesso*	迴廊*wui4 long4* 'corridor'
Archives (1)	*Arquivo*	檔案館*dong2 on3 gun2* 'Archives'
Bronze statue (1)	*Monumento* 'Monument'	銅 像 *tung4 zoeng6* 'bronze statue'
Convention & Entertainment Centre (1)	*Centro de Convenções e Entretenimento*	中心*zung1 sam1* 'centre'
Cultural Centre (1)	*Centro Cultural*	中 心 *zung1sam1* 'centre'
Cultural Village (1)	*Aldeia Cultural*	村*cyun1* 'village'
Ecumenical Centre (1)	*Centro Ecuménico*	苑*jyun2* 'garden'
Gate (1)	*Portas*	關閘 *gwaan1 zaap6* 'border gate'
Mosque and Cemetery (1)	*Mesquita e Cemitério*	寺 *zi6*, 墳 場 *fan4 coeng4*'temple, cemetery'
Pavilion (1)	*Pavilhão*	熊貓館 *hung4 maau1 gun2* 'panda house', 動物 館 *dung6 mat6 gun2* 'zoo'
Square (1)	*A Praça* 'The Square'	廣場 *gwong2 coeng4* 'wide square'
Wharf (1)	*Doca*	碼頭*maa5 tau4* 'pier'

Numbers in brackets report the occurrences of the attested forms. Entries appear with numbers of tokens and in alphabetical order.

two different periods (433,574 attested entries in July 6 2022, 455,383 entries in December 7 2023). We then analysed their spelling and their linguistic properties, and their geographic distribution to individuate their dialectal roots. We also incorporated the temporal dimension to observe trends in the growth of linguistic data in OSM and pinpoint the specific locations where such growth occurs. We can thus discuss three key novel results (N.B. We present the data for the figures in the Appendix, while all the other relevant data were available in Supplementary file B, C).

First, OSM offers a higher number of toponyms than some AGI sources, at least with respect to urban toponyms. We addressed this aspect by extracting toponyms from the YellowPages online directory.[5] A clarification on this directory as an AGI source is necessary, before we proceed. The YellowPages directory has commercial purposes, since it offers addresses and locations of various commercial activities that are willing to buy this service. Thus, the lists of places with commercial functions are not necessarily exhaustive. However, the maps and place directories are based on official gazetteers provided by local and national administrations. Crucially for our purposes, the YellowPages directory includes urban toponyms, and it includes data of AGI origin not offered by volunteers. It thus approximates an AGI source.[6]

We obtained 213,218 toponyms, i.e. less than half of the toponyms extracted via OSM. Crucially, the YellowPages directory includes directories for minor urban centres, villages, and hamlets. However, these province-specific directories tend to offer lower-resolution maps than those for major urban centres (e.g. 1:5000 against 1:3000, and against OSM's 1:1000 scale). They can overlook villages and hamlets that may be too small to appear at these resolutions, and therefore report lower numbers of toponyms. Upon calculating the Jaccard index for these lists (score: 0.0088), we also confirmed that the YellowPages directory only includes a part of the toponyms found in OSM. We can thus conclude that OSM can currently offer a higher quantity of toponyms and a better coverage of their distribution on the Italian territory than the AGI source YellowPages.

Second, the geographical distribution of urban toponyms, irrespective of their linguistic origins, appears heterogeneous. However, at a regional scale, a more nuanced picture emerges that involves different types of homogeneous distributions. Figure 7a, b offer the distributions in terms of tokens and percentages associated with each region in the two time-spans under investigation. We have retrieved data in two different periods to detect the dynamic evolution of the source. This evolution (i.e. increase of instances) appears in the lower panel of Fig. 7, again in terms of tokens and percentages:

Given the size and relevance of these updates, we can infer that most contributors concentrated on these regions to remove perceived or real gaps in toponyms' coverage. A further finding is that some regions include urban centres that cover most toponyms, and thus indirectly determine heterogeneous distributions. Other regions, instead, feature more spatially homogeneous distributions of toponyms. This is the case irrespective of the number and size of urban centres. We can analyse this pattern via quantitative and qualitative insights based on the most recent dataset as a reference (December 7, 2023). The quantitative insights are as follows. The impact of the most prominent urban centre within each region, automatically retrieved by inputting the name in Italian, appears in Fig. 8a.

The Northern region of Lombardy comprises 11.56% of the total toponyms; its administrative capital and global economic hub, Milan, covered 8.96% (4715 tokens) of the total. Central regions Lazio and Molise respectively comprised 8.20% and 2.98% of the total, instead. Lazio includes the national capital city, Rome (45.09% of the total; 16843 tokens). Molise is a small region to the West of Lazio, with Campobasso being its most important urban centre. The dark blue colour for Molise indicates that Campobasso includes the majority of this region's tokens. The light blue colour for Lombardy indicates that Milan does not include most tokens. Lazio's shade represents an intermediate to high concentration example. More in general, most Italian regions feature homogeneous distributions of toponyms (i.e. no urban centre covers most tokens: lighter blue shades). However, regions with heterogeneous distributions also occur (darker blue shades).

The qualitative insights are as follows. Basilicata is a mostly mountainous region including national parks and a few minor cities, in the South of the country (dark blue shade in Fig. 8a). Its administrative seat, Matera and the other urban centres appear scattered on this territory. These centres correspond to clusters of

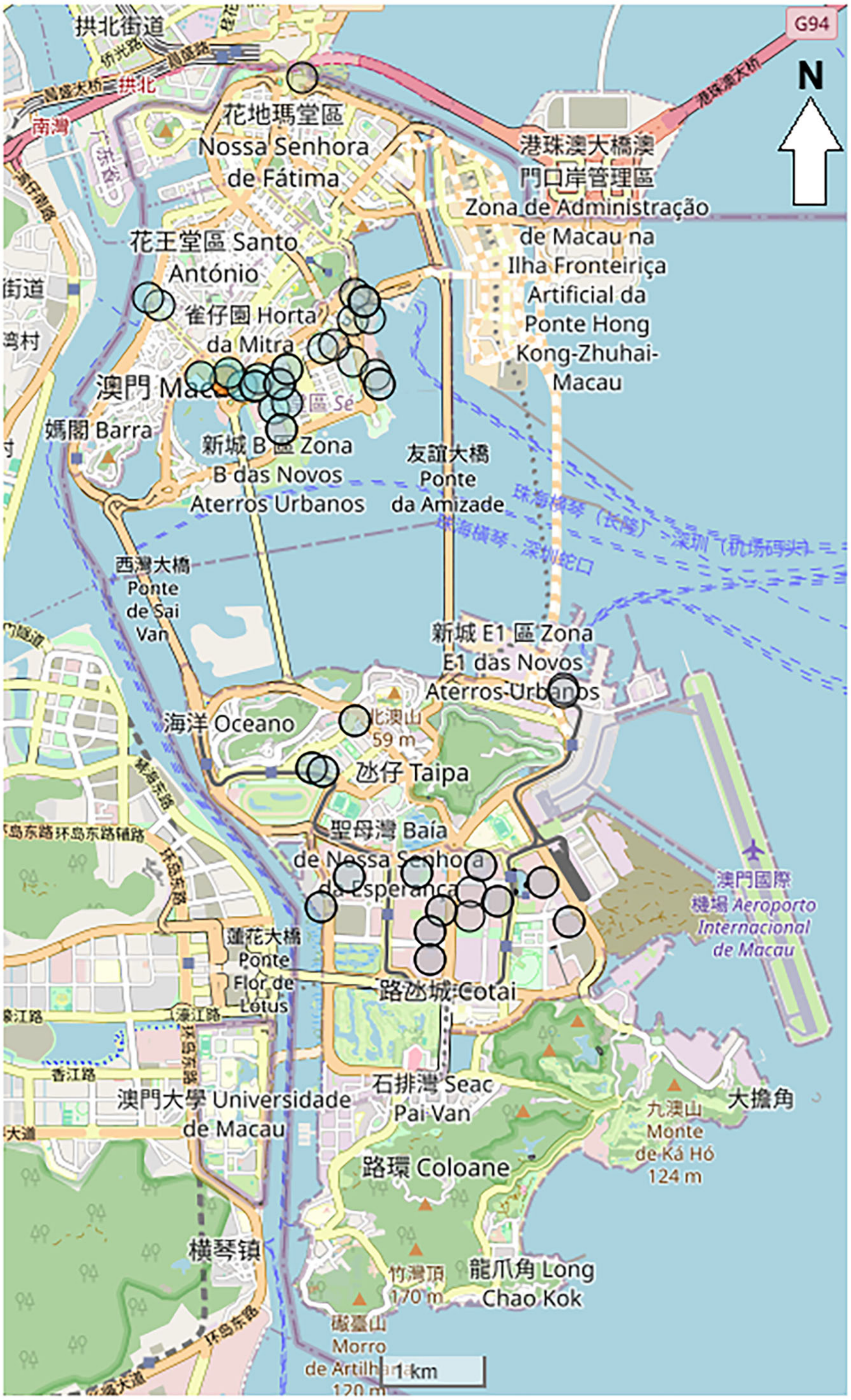


Fig. 6 Map of Casinos. As we explain in the main text, casinos mostly lie in the more urbanised parts of the city. Coloane lacks any of these places, being mostly green places[4].

urban toponyms in spaces mostly devoid of these toponyms (light yellow shade in Fig. 8b). Liguria, on the other hand, is the coastal region in the North-West of the country, and offers a clear example of urbanisation in a limited space. That is, the region has a wealth of small urban centres and only one centre covers a relative majority of toponyms, its administrative seat, Genoa. However, the distribution of these centres appears (relatively) dense and evenly distributed.

We present these patterns in Fig. 8b, in which we plotted a ratio calculated by the number of toponyms divided by the number of km^2 of surface.[8] Aggregating all regions, we observe a strong correlation between the number of tokens and the size of the surface in km^2 (Pearson's $r = 0.82$, $p < 0.01$). These data therefore show that distribution may vary from a homogeneous to a heterogeneous pole. Italian regions generally feature many urban centres, as Italy has a long history of diffused and pervasive urbanisation. In some regions, however, historical matters caused few cities, usually the regions' administrative seats, to develop a higher platial and toponymic relevance (Ursini 2020). Such differences indirectly appear in OSM as geographical differences in toponyms distribution.

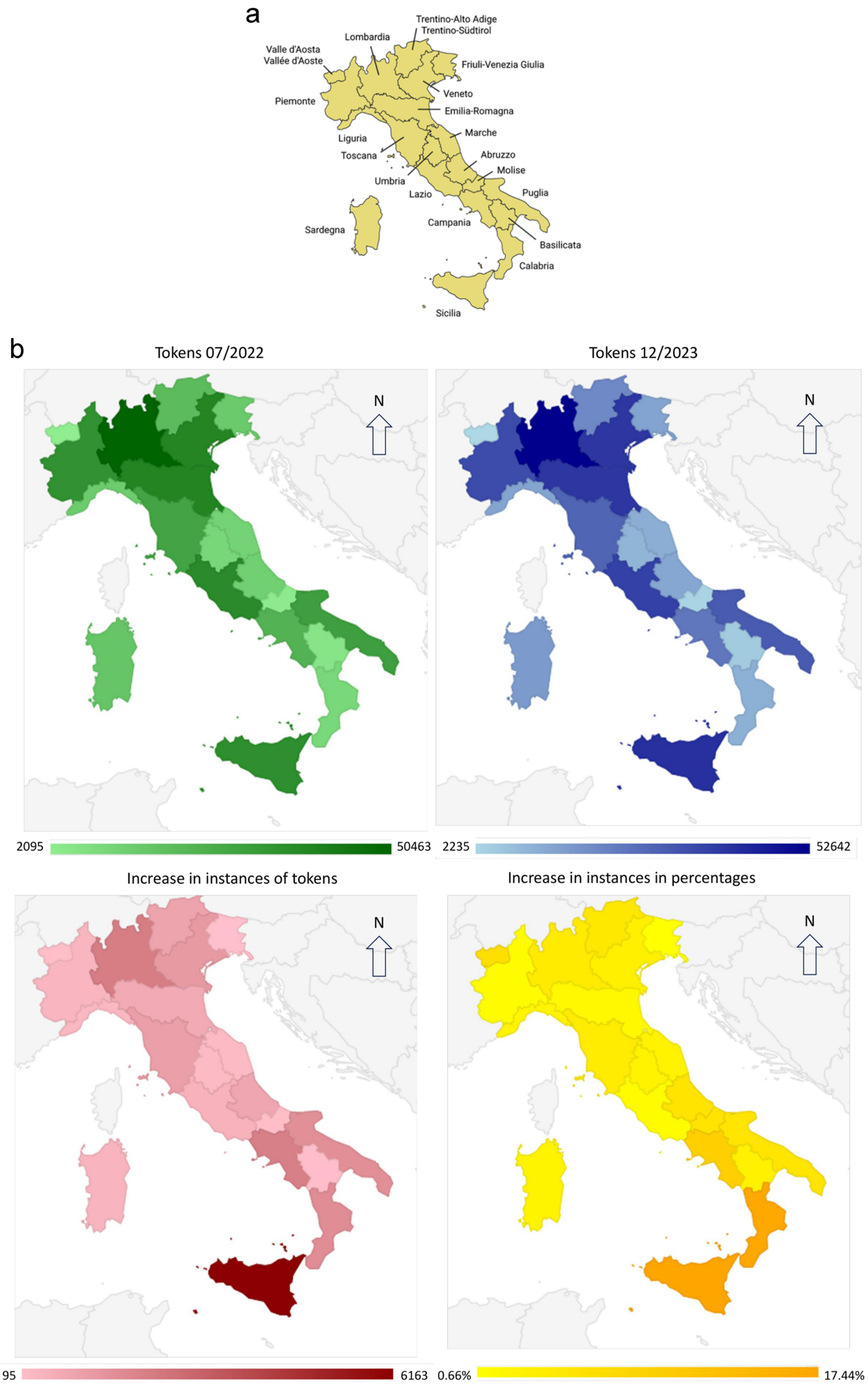


Fig. 7 Distribution (in tokens) and increase of instances (in tokens and percentages) in the target time span. The figure in panel **a** shows that show that some regions include the highest percentages of toponyms due to their size and the higher number of urban centres (e.g. Lombardy in the North, Lazio in the centre, Sicily in the South). The figure in panel **b** also shows that also show that some regions (e.g. Lombardy, Sicily) also involved updates of considerable in the target time span. We use the figure in panel **a** also to introduce the names of the 20 administrative regions.[7].

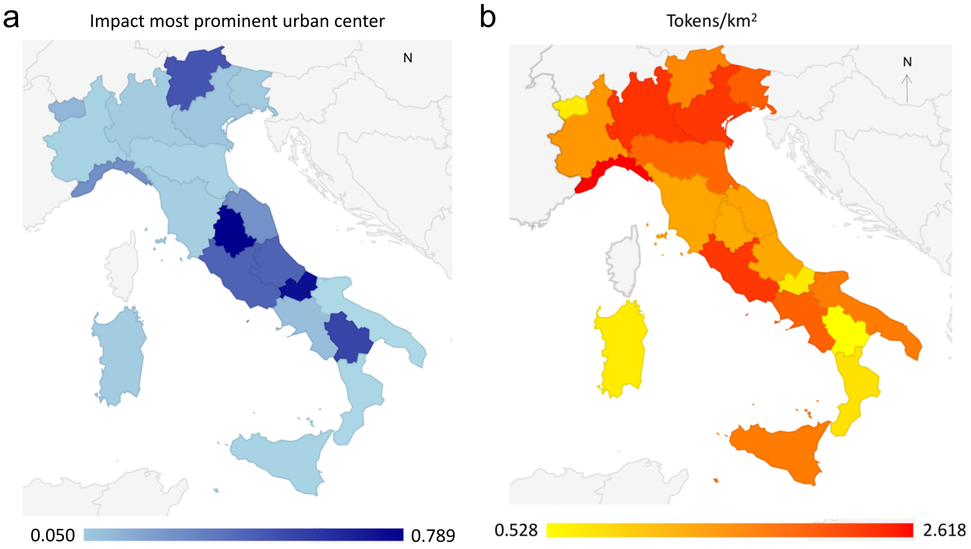


Fig. 8 Distribution of geocodes. a Regions with administrative seats including the highest number of geocodes are in dark blue. These regions are: Liguria, Trentino-Alto Adige (North); Umbria, Lazio, Abruzzo, Molise (Centre); Basilicata (South). **b** Distribution of tokens/surface ratio. Lighter-shaded regions have low numbers of tokens per square kilometre; darker-shaded have higher numbers. Lighter-shaded regions are Valle D'Aosta (North); Sardinia (Island, Centre), Molise (Centre); Basilicata, Calabria (South).

Third, linguistic analysis and geographical distribution interact in at least one aspect. In the original study, we found that dialectal urban toponyms correlate with the geographical distribution of their respective dialects. For instance, one can find toponyms in the Neapolitan dialect in Naples, the Campania region in which Naples is located, and in neighbouring regions in which this dialect finds currency (e.g. Molise). In this study, we analysed the distribution and density of dialectal toponyms in each of the 20 administrative regions, and assessed which toponyms preserved their original dialectal spelling. Our study-specific goal was to analyse whether the preservation of these roots is correlated to the geographical distribution and linguistic properties of these toponyms. We verified whether dialectal toponyms would crop up in highly delimited places (e.g. single cities), and whether their spelling was compatible with standard Italian spelling. The results of this update are in Fig. 9.

As the figure shows, the distribution of dialectal urban toponyms tends to correlate with specific cities and to express local dialectal identities. For instance, one can find toponyms including terms *rua* in the city of Ascoli Piceno (region, Marche), *crosa* in Genoa (Liguria) and *venula* in Erice (Sicily). However, certain terms (e.g. *calle*) can also occur outside the cities in which they have the densest distribution; nevertheless, as our figures show, they are exceptions to this rule. Similarly, most terms and toponyms feature spelling forms that are adapted to Italian rules, e.g. again *crosa* for Genoese/Zeneize *creusa*. Toponyms preserving the original spelling are thus rare but also homogeneously scattered across regions: each Italian region contributes dialectal toponyms to Italian.

In the original study, we also found the presence of toponyms from minority languages at a regional level. For instance, overpass reports German toponyms in Sud-Tyrol/Alto Adige, a province with an ample German-speaking population (e.g. *Garibaldistraße* 'Garibaldi Street'), and other toponyms belonging to other local linguistic minorities (e.g. Sardinian, Friulan: Samo and Ursini 2023). In this update, we offer a geographical analysis of the distribution of these minority languages and their toponyms. This pattern emerges once we analyse the number of generic terms that can be associated with each region. For this purpose, we make use of a Type/Token ratio (TTR) to detect the nature of generic terms (generic terms=types). A smaller TTR means more standardised forms, whereas a higher TTR means more forms pro-token. We can observe the geographical distribution of TTR's in Fig. 10.

As the figure shows, Valle D'Aosta (North-West) and Trentino/Alto Adige (Northeast) regions emerge as bilingual regions: French (Valle D'Aosta) and German (Trentino/Alto Adige) co-exist with Italian. Most generic terms in the analysis belong to these two languages. In the case of the Alto Adige province, the role of German stems from the fact that this region was part of the German-speaking portions of the Austro-Hungarian empire until WWI. Furthermore, toponyms belonging to these languages mostly occur in urban places such as Bozen/Bolzano, the administrative seat of the Alto Adige province. Toponyms from minority languages thus strongly tend to correlate in distribution with the regions and cities in which their respective languages have official status. Their distribution appears heterogeneous on the national territory because languages find use in specific regions. With these three results at our disposal, we can answer the first two research questions as follows.

Regarding **RQ1**, OSM offers coverage of urban toponyms for the Italian territory that is now considerably superior to some AGI sources. For this study, the density of toponyms may reflect the homogeneous distribution of places in Italian urban centres, but also contributors' commitment to pursue a spatially homogeneous coverage of toponyms. Heterogeneous distributions certainly emerge, but they seem to reflect two factors. First, in some regions, few urban centres exist that cover most toponyms in their region. Second, coverage may still be incomplete: contributors may simply temporarily focus on regions with sparser toponyms, for manifold contingent reasons. Let us note that in this study, we did not find toponyms for places attested at finer-grained resolution levels (e.g. public gardens), as in the case of the first study. We believe that the possibility of finding such toponyms in future updates for OSM may address this issue.

Regarding **RQ2**, if we consider Italian, the dialects of Italian and the minority languages as expressing a form of multi-lingualism, then we can conclude that OSM indirectly captures this multi-lingualism. However, AGI sources (here, the Yellow-Pages gazetteer) certainly report toponyms in multi-lingual contexts: OSM seems to perform at equal levels of coverage.

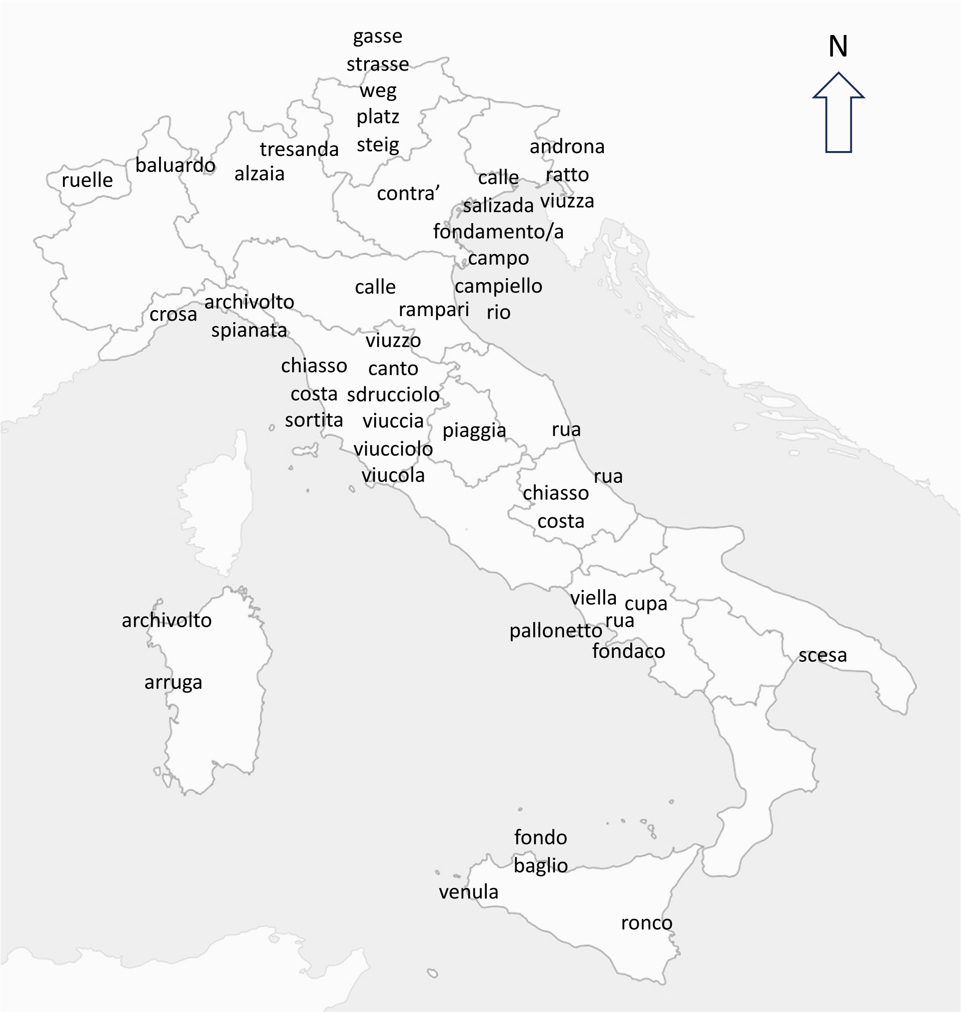


Fig. 9 Distribution of 'local' generic terms across the peninsula. Terms appear on the surface of the region in which they occur in local toponyms. For instance, *chiasso* appears in toponyms from Tuscany and Abruzzo; we thus mapped the term onto these two regions, in the map.

Asymmetries and forms of heterogeneity in OSM occur in two cases. First, contributors must still add missing data for more sparsely populated and less economically relevant regions (e.g. Basilicata and Molise, at the time of writing). Second, contributors seldom insert dialectal toponyms that lack such an official status (e.g. dialectal names for Sicilian villages). Such data would count as bottom-up contributions that would prove the superior empirical coverage of OSM as a VGI source. Nevertheless, since OSM operates at a higher scale of resolution than some AGI sources, it apparently offers a more homogeneous picture of multi-lingual data.

Discussion

We believe that four possible generalisations emerge from our data: one for each research question and one overarching generalisation emerging from the meta-study itself.

First, **RQ1** addresses the issue of toponyms and their numbers in OSM regarding their heterogeneous geographical distribution. Our novel data show that OSM often includes different potentially higher amounts of toponyms than some AGI sources. However, OSM maps operate at relatively high levels of resolution (again, 1:1000 for city environments: Curran et al., 2012; Mooney et al., 2017). The selected AGI sources implement higher or lower resolutions, instead (e.g. respectively the Macao APP at 1:500; the Italian YellowPages at 1:3000). Our novel meta-study finding is that OSM’s structural properties may determine contingent heterogeneity patterns. Contributors may operate at different resolution levels as they insert information according to their specific and current knowledge (cf. also Hecht et al., 2013; Arsanjani et al., 2015; Antoniou and Skopeliti, 2017; Senaratne et al., 2017; *OpenStreetMap Wiki—Map features*, n.d).

A conclusion that we can draw from these results is that the quality of OSM data depends on two factors. The first is the quality of contributors’ work. The second is the quality of the information and sources that contributors can implement in this work (e.g. Goodchild, 2007; Keßler et al., 2009; Sui and Goodchild, 2011; Schäfer and Kieslinger, 2016; Senaratne et al., 2017; Novack et al., 2022). By answering **RQ1**, we therefore show that OSM is a valuable source of information about toponyms for GIS and linguistic disciplines, if researchers perform careful triangulation with other AGI sources.

Second, **RQ2** addresses the issue of coverage asymmetries regarding toponyms in multi-lingual regions; it asks where and at what scales these symmetries arise. Our novel findings suggest that multi-lingual coverage asymmetries may be due to historical reasons that can emerge only once researchers perform a careful linguistic analysis (cf. Blair and Tent, 2015, 2021; Qian et al., 2016). We can thus conclude that spatial asymmetries are indirect evidence of languages’ temporal, socio-historical roles in cities (e.g.

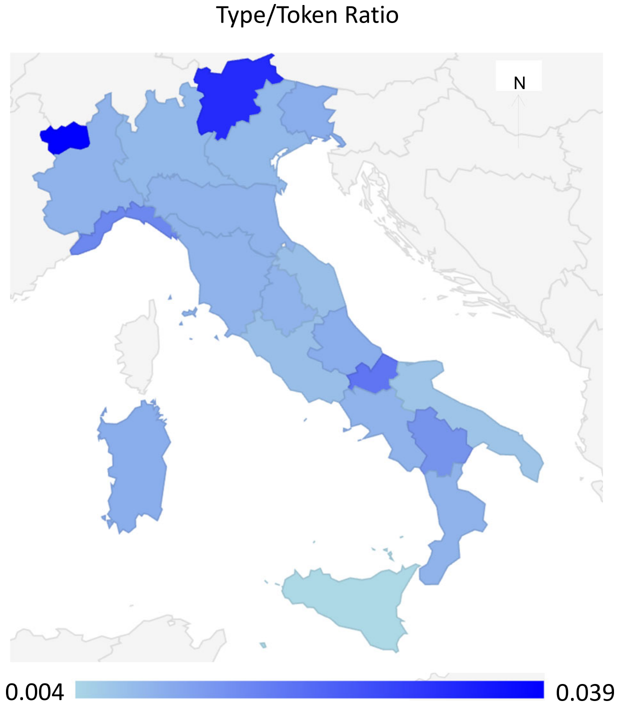


Fig. 10 Geographical distribution of TTR of generic terms from minority languages. The dark blue regions are Valle D'Aosta (North-West) and Trentino-Alto Adige (Northeast). The minimal value '0.004' means that 4 out of 1000 tokens were tokens for generic terms originating in minority languages. The maximal value '0.039' means that we found 39 out of 1000 tokens.

Macao), regions and nations (e.g. Italy), and mirror these roles in local toponyms (Stolz and Warnke 2018; Cavallaro et al., 2019).

Another important observation is that our novel data show that asymmetries may not only originate in cases of incomplete coverage. Once more, contributors working on less developed urban zones or belonging to less privileged communities and lacking access to AGI sources may face deep challenges in working on OSM updates (Antoniou and Skopeliti, 2017; Bortolini and Camboim, 2019; Carraro, 2021; Romm and McKenzie, 2023; Seto, 2022; Touya et al., 2017). However, OSM contributors can overcome such coverage asymmetries via systematic, collective, and continuous updates. By answering **RQ2**, we therefore show that linguistic disciplines can benefit from OSM as a multilingual information source. Again, careful triangulation with AGI sources appears as a methodological necessity.

Third, **RQ3** addresses the issue of how these results can inform GIS research, toponomastic, and broader linguistic research focusing on toponyms. From our answers to **RQ1** and **RQ2**, we can offer the following general answer. Regarding GIS, we offer further evidence that OSM is evolving into a multi-source system, although its VGI roots remain evident (Bravo and Sluter, 2022; Hu et al., 2022). We also offer further evidence that toponym retrieval and analysis from OSM can benefit GIS disciplines. If researchers extract toponyms from a given region, then they can analyse the information encapsulated in tags including these toponyms for any other purposes (Daniel and Mátyás, 2022; Hall and Jones, 2022; Qian et al., 2016).

Regarding linguistics and sub-disciplines such as toponomastics, socio- and geo-linguistics, we also offer novel proof that OSM can offer quantitatively robust data sets involving multilingual contexts. Other works that have implemented the methodology outlined in this work have instead offered corroborating evidence regarding the use of these data for theoretical (e.g. grammatical, lexical) analyses (e.g. Chinese, Portuguese, German, Italian, Italian dialects, French: Xie et al. 2023, Samo and Ursini 2022, 2023, 2024). Again, such results are reliable if researchers triangulate OSM data with AGI data.

Fourth, we can offer three observations on how triangulated OSM data can play a novel role in other disciplines that, however, have a more speculative nature. These observations can also act as potential starting points for future research, we believe. Therefore, we discuss them in some detail, before moving to the conclusions.

A first observation is that the digitisation of historical gazetteers has recently witnessed a considerable growth and has informed historical toponomastics (Southall and Aucott, 2019; Grossner et al., 2022). OSM as a VGI source may eventually come to include a historical, diachronic component to the representation of toponyms (e.g. information about the Colosseum during the Imperial Age of Rome). For the time being, however, OSM seems focused on offering synchronic maps of the world. One should wait for future developments regarding how this platform can include 'maps of the past', before one can plan and carry out future research projects.

A second observation is that in psychology and cognitively oriented GIS works, researchers have analysed OSM tags as mental models or cognitive maps of places. They thus have used toponyms for information extraction regarding these models (Mayer et al., 2022; Mocnik, 2022; Ursini and Zhang 2023). These and other works have thus suggested that descriptively rich OSM tags for places can approximate standard conceptions of the multi-dimensional, relational notion of place suggested in the literature (e.g. Cresswell, 2014; Malpas, 2018; Tuan, 1977). Future psycho-linguistic and cognitively oriented GIS works may therefore potentially explore OSM data as shedding light on the semantic content of toponyms.

A third observation involves critical toponymy and other geographic disciplines focusing on toponyms. Most works in this discipline analyse the dynamics of power underpinning place naming practices and toponyms' representation in maps (e.g. Azaryahu 2011; Rose-Redwood et al., 2010, 2018). A central concern is thus to analyse how different groups can achieve representation in the urban landscape, and how they can find ways for their (urban) toponyms to appear in gazetteers (e.g. Alderman, 2022; Basik, 2020; Cavallaro et al., 2019; Walkowiak 2024). We believe that, at a methodological level, our study shows that OSM can potentially provide important data for this type of research, as we also outlined in our **RQ2** answer. Critical toponymy studies may, for instance analyse how power conflicts in 'place name reporting' occurs within OSM communities, and how these conflicts can affect the Open Access nature of the project. Future works in this line of research may therefore benefit from the open access philosophy of this source, even though careful analysis of OSM's specific issues of accessibility becomes necessary (Bortolini and Camboim, 2019; Carraro, 2021).

Overall, by answering **RQ3**, we offer further evidence supporting previous findings in the GIS literature. Such evidence also offers novel insights on the role of OSM in research on toponyms. We must, however, stress, once more, that one must triangulate OSM data with data from AGI sources. OSM as a VGI source and AGI sources can provide maps of the world and its places that may involve various forms of missing data, due to manifold contingent reasons. OSM contributors may certainly face external challenges and pressures, but professional compilers of AGI sources may also lack access to toponymic data, even if in a less pronounced manner. Careful triangulation of sources becomes therefore necessary, when studying toponyms data (Hu, Al-Olimat, et al., 2022). Our paper offers novel proof supporting this methodological claim, and a methodology that may find implementations in any context of analysis.

Conclusions

The goal of this paper has been to offer an analysis of OSM as a research tool for linguistics and GIS disciplines. The paper has therefore offered an answer to three research questions that arise once we address the role of OSM as a rich but heterogeneous VGI source for studies on toponyms. These three questions focus on the quantity and accuracy of toponyms' coverage in OSM and how this coverage compares with AGI sources (**RQ1**), what are the possible asymmetries attested on regions involving multilingual realities and toponyms (**RQ2**), and how GIS and linguistic studies can profit from studying these problems (**RQ3**). The paper has thus introduced a novel methodology operating at a multi-scalar (city, provincial, region and national level) and multi-lingual (Portuguese, Chinese and Italian languages and dialects) level. The paper has used these results to answer these research questions in detail, and to show that research on toponyms across different disciplines and cultural/linguistic contexts can amply benefit from data originating in this source.

Our consequent overarching answer, then, is that OSM offers tools for the processing of toponyms' information and the management of this result that are comparable to AGI sources. However, regions and national territories are complex types of places that may present asymmetries in places' distribution. Such asymmetries emerge in OSM and any other source as heterogeneous toponyms' distributions. Therefore, a fuller multi-source approach, constantly comparing and integrating OSM data with AGI sources, is necessary to guarantee the integrity of these processes. If one meets these conditions, OSM can offer an increasingly sound platform for cross-cultural inter-disciplinary research on toponyms and places in GIS, linguistics, and other sciences. For further endeavours, we must, however, wait for future research.

Data availability

The datasets generated during and/or analysed during the current study are available in the OSF repository, https://osf.io/fw37d/.

Received: 21 October 2024; Accepted: 12 May 2025;
Published online: 11 June 2025

Note

[1] The link is available at: https://openstreetmap.com, last access 12.11.2024.

[2] The link is available at: https://overpassturbo.eu/, last access 12.11.2024.

[3] The link is available at https://webmap.gis.gov.mo/InetGIS/eng/index.html, last access 12.11.2024.

[4] The screenshots are from OSM. The link is available at: https://osmand.net/map/#13/22.1743/113.5612, last access 12.11.2024.

[5] The link is available at: https://paginegialle.it, last access 12.11.2024.

[6] The existence of commercial geographic services opens the possibility of considering 'pure' AGI and VGI sources as poles of a conceptual continuum. Please consult Sarkar and Anderson, (2022); Westerholt. (2019) for further discussion.

[7] The link is available at: https://it.wikipedia.org/wiki/File:Regions_of_Italy_with_official_names.png, last access 12.11.2024.

[8] Data retrieved from the website of the National Institute of Statistics of Italy: https://www.istat.it/it/archivio/137001, last access 12.11.2024.

References

Ahmadian S, Pahlavani P (2022) Semantic integration of OpenStreetMap and CityGML with formal concept analysis. Trans GIS 26(8):3349–3373

Ahmed MdK (2022) Converting OpenStreetMap (OSM) data to functional road networks for downstream applications. Preprint at https://doi.org/10.48550/arXiv.2211.12996

Alderman MH (2022) Commemorative place naming: to name place, to claim the past, to repair futures. In: Giraut F, Hossay-Holzschuch M (eds) The politics of place naming: naming the world. John Wiley and Sons, pp 29–46

Almendros-Jiménez JM et al. (2021) Metamorphic testing of OpenStreetMap. Inf and Sof Tec 138(1):106–319

Anderson J, Sarkar D (2020) Curious cases of corporations in OSM. In: Anderson J, Sarkar D, Minghini M (eds) Proceedings of the academic track, state of the map,

Antoniou V, Skopeliti A (2017) The impact of the contribution microenvironment on data quality: the case of OSM. In: Foody, G et al. (eds) Mapping and the citizen sensor. Ubiquity Press, London, pp 165–196

Antoniou V, Touya G, Raimond AM Quality analysis of the Parisian OSM toponyms evolution. In: Capineri C, Haklay M, et al. (eds) European handbook of crowdsourced geographic information. Ubiquity Press, London, pp 97–112

Arsanjani JJ, Zipf A et al. (eds) OpenStreetMap in GIScience: experiences, research and applications. Springer, Berlin

Azaryahu M (2011) The critical turn and beyond: the case of commemorative street naming. ACME Int J Cri Geo 10(1):28–33

Basik S (2020) Rethinking the toponymic politics in Belarus in the 20-21 centuries. Toward the post-colonial perspective. J Geo Polit Soc 10(3):5–15

Basik S (ed) (2021) Urban place names: Special issue. Urb Sci 80(4):1–121

Berruto G (2012) Sociolinguistica dell'Italiano Contemporaneo, 2nd ed. Carocci Editore, Rome

Blair D, Tent J (2021) A revised typology of place names. Names 69(4):1–15

Blair D (2015) Tent J feature terms for Australian toponymy. ANPS Technical Paper 3. https://www.anps.org.au/upload/ANPSTechPaper3.pdf

Bossong G (2016) Classifications. In: Ledgeway A, Maiden M (eds) The Oxford guide to the romance languages. Oxford Academic, Oxford

Botha W, Moody A (2021) English in Macau. In: Botha W, Bolton K, Kirkpatrick A (eds) The handbook of Asian englishes. John Wiley and Sons, New York, pp 3–30

Bravo JVM, R Sluter CR (2022) Crowdsourcing map-using and map-generating tasks into OpenStreetMap. Pro Geo 2(2):248–262

Calafiore G (1975) Termini geografici dialettali in Italia. Istituto di Geografia, Firenze

Carraro V (2021) Jerusalem online: critical cartography for the digital age. Springerlink, Berlin

Cartography and Cadastre Bureau of Macau SAR (2021) 澳門特別行政區數碼化地圖唯讀光碟A類/CDROM de Carta-base (Tipo A) da Regi˜ao Administrativa Especial de Macau. [CD-ROM of the paper versi´on (type A) of the Special Administrative Region of Macau]'. In:澳門特別行政區政 府地圖繪製暨地籍/Macau: Direção dos Serviços de Cartografia e Cadastro

Cassi L (2015) Nomi e Carte: Sulla toponomastica della Toscana. Pacini Editore, Firenze

Cassi L, Marcaccini P (1998) Toponomastica, beni culturali e ambientali. Gli indicatori geografici per un loro censimento. In: NN. AA. (eds.) Collezioni Memorie della Società Geografica Italiana pp 655–1097

Cavallaro F, Perono Cacciafoco F, Xuan Tan Z (2019) Sequent occupance and toponymy in Singapore: the diachronic and synchronic development of urban place names. Urb Sci 3(3):77–98

Cerri M et al. (2021) Are OpenStreetMap building data useful for flood vulnerability modelling? Nat Haz and Ear Sys Sci 21(2):643–662

Chiappinelli L (2013) Nomi di Luogo in Campania. Percorsi Storico-Etimologici. Edizioni Scientifiche Italiane, Roma

Coates R (2007) Urban toponymy. Ono 42:1–235

Cresswell T (2014) Place: an introduction. John Wiley & Sons, New York

Curran K, Crumlish J, Fisher G (2012) OpenStreetMap. Int J Int Com Sys Tec 2(1):69–78

Curran K, Crumlish J, Fisher G (2013) OpenStreetMap. In: Curran K, Crumlish J, Fisher G (eds) Geographic information systems: concepts, methodologies, tools, and applications. IGI Global, pp 540–549

Damico J, Tetnowski J (2014) Triangulation. In: Forsyth C, Copes H (eds) Encyclopedia of social deviances. Sage Publications, Riverside, pp 709–721

Daniel N, Màtyàs G (2022) Citizen science characterization of meanings of toponyms of Kenya: a shared heritage. GeoJ 88(3):1–22

David J (2011) Commemorative place names – their specificity and problems. Names 59(4):214–228

De Mauro T Il (2020) Dizionario della Lingua Italiana. Paravia Editrice, Milano

Everton Bortolini E, Camboim SP (2019) Mapeamento colaborativo de favelas com a plataforma openstreetmap collaborative slum mapping with Openstreetmap. In: Colombo VP, Bassani J, Torricelli GP, de Araújo SA (eds) Mapeamento participativo: tecnologia e cidadania. Editora da Faculdade de Arquitetura e Urbanismo da Universidade de São Paulo, São Paulo, pp 33–52

Fasold R, Connor-Linton J (2014) Introduction. In: Fasold R, Connor-Linton J (eds) An introduction to language and linguistics, 2nd edn. Cambridge Univ. Press, Cambridge, pp 1–21

Fize J, Moncla L, Martins B (2021) Deep learning for toponym resolution: geocoding based on pairs of toponyms. Int J Geo Inf 10(12):800–818

Fotheringham AS, Wilson JP (2007) Geographic information science: an introduction. In: Fotheringham AS, Wilson JP (eds) The handbook of geographic information science, Wiley, New York, pp 1–7

Gamerschlag T et al. (eds) (2015) Meaning, frames, and conceptual representation. 2. Walter de Gruyter GmbH & Co KG, Berlin

Garba S et al. (2022) Quality analysis of OpenStreetMap and digital elevation data based North-Western Nigeria. FIG congress 2022 volunteering for the future - geospatial excellence for a better living Warsaw, Poland

Gnatiuk O, Melnychuk A (2020) Geopolitics of geographical urbanonyms. A Uni car. Geo 33(2):255–268

Goodchild MD (2007) Citizens as sensors: the world of volunteered geography. GeoJ 69(4):211–221

Goodchild MD (2011) The convergence of GIS and social media: challenges for GIScience. Int J Geo Inf Sci 25(11):1737–1748

Grossner K, Grunewald S, Mostern R (2022) Bringing places from the distant past to the present: a report on the World Historical Gazetteer. Int J Dig Lib 24(1):1–4

Hall MM, Jones CB (2022) Generating geographical location descriptions with spatial templates: a salient toponym driven approach. Int J Geo Inf Sci 36(1):55–85

Hecht R, Kunze C, Hahmann S (2013) Measuring completeness of building footprints in OpenStreetMap over space and time. ISPRS Int J Geo-Inf 2(4):1066–1091

Holthaus T, Thiemermann A (2022) Identifikation deutscher Strasenentwurfsklassen im Strasennetz von OpenStreetMap. Roa Net 8(1):93–105

Hu X, Al-Olimat HS et al. (2022) GazPNE: annotation-free deep learning for place name extraction from microblogs leveraging gazetteer and synthetic data by rules. Int J Geo Inf Sci 3(2):310–337

Hu X, Zhou Z et al. (2023) Location reference recognition from texts: a survey and comparison. ACM Comput Surv 56(6):1–37. https://doi.org/10.1145/3625819

Jaccard P (1901) ´Etude comparative de la distribution florale dans une portion des Alpes et des Jura. Bul de la Soc vau sci nat 37(4):547–579

Jaljolie R et al. (2023) Evaluating current ethical values of OpenStreetMap using value sensitive design. Geo-spa Inf Sc 26(3):362–378

Keßler C (2017) OpenStreetMap. Encyclopedia of GIS. pp1493–1498

Keßler C, Janowicz K, Bishr M (2009) An agenda for the next generation gazetteer: geographic information contribution and retrieval. In: Proceedings of the 17th ACM SIGSPATIAL international conference on advances in Geographic Information Systems. Boston, pp 91–100

Lorenz M, Aisch G, Kokkelink D (2012) Datawrapper: create charts and maps [software]. https://www.datawrapper.de

Machado AA et al. (2021) Informação geográfica voluntária: o potencial das ferramentas colaborativas para a aquisição de nomes geográficos. Rev bra de Geo 66(2):239–253

Malpas J (2018) Place and experience: a philosophical topography. Routledge, London

Marcato C (2009) Nomi di Persona, nomi di luogo. Il Mulino, Bologna

Mastrelli CA (2005) La normativa sull'onomastica e gli stradari. In: Mastrelli CA (ed) Odonomastica. Atti del convegno, 2002 Trento Provincia autonoma, Soprintendenza per i beni librari e archivistici, Trento

Mastrelli CA (2005) La normativa sull'onomastica e gli stradari}. In: Mastrelli CA (ed), Odonomastica. Criteri e normative sulle denominazioni stradali. Atti del convegno (Trento, 25 settembre 2002). Provincia autonoma, Soprintendenza per i beni librari e archivistici, Trento

Mayer M, Heck DW, Mocnik F-B (2022) Using OpenStreetMap as a data source in psychology and the social sciences. Preprint at PsyArXiv https://doi.org/10.31234/osf.io/h3npa

Mocnik F-B (2022) Putting geographical information science in place–towards theories of platial information and platial information systems. Pro Hum Geo 46(3):798–828

Mooney P, Juhász L (2020) Mapping COVID-19: how web-based maps contribute to the infodemic. Dia Hum Geo 10(2):265–270

Mooney P, Grinberger AY et al. (2021). OpenStreetMap data use cases during the early months of the COVID-19 pandemic. In: Rajabifard A, Paez D, Foliente G (eds) Covid-19 pandemic, geospatial information, and community resilience: global applications and lessons. CRC Press, Boca Raton, pp 171–186

Mooney P, Minghini M et al. (2017) A review of OpenStreetMap data. In: Foody G, et al. (eds) Mapping and the citizen sensor. Ubiquity Press, London, pp. 37–59

Nasr Naim E et al. (2023) Exploring spatio-temporal patterns of OpenStreetMap (OSM) contributions in heterogeneous urban areas. Bol Ciénc Geod 29:e2023005

Novack T, Vorbeck L, Zipf A (2022) An investigation of the temporality of OpenStreetMap data contribution activities. Geo-Spa Inf Sci 27(2):259276

Perdana AP, Ostermann O (2018) A citizen science approach for collecting toponyms. ISPRS Int J Geo-Inf 7(6):214–222

Perono Cacciafoco F, Cavallaro F (2023) Place names: approaches and perspectives in toponymy and toponomastics. Cambridge University Press, Cambridge

Qian S, Kang M, Wang M (2016) An analysis of spatial patterns of toponyms in Guangdong, China. J Cult Geo 33(2):161–180

Queirazza GG et al. (eds) (1990) Dizionario di Toponomastica Italiana. Utet, Milano

Rajšp A, Hericko M, Fister I Jr (2021) Preprocessing of roads in OpenStreetMap based geographic data on a property graph. In: VV.AA. (eds) Central European conference on information and intelligent systems. Faculty of Organization and Informatics Varazdin, pp 193–199

Romm D, McKenzie G (2023) Platial Rhythm. In: VV.AA. (eds) Spatial knowledge & information. Springer, Berlin, p. 1–9

Rose-Redwood R, Alderman DH, Azaryahu M (2010) Geographies of toponymic inscription: new directions in critical place-name studies. Pro Hum Geo 34(4):453–470

Rose-Redwood R, Alderman DH, Azaryahu M (2018) The political life of the urban streetscape: naming. Politics and Place. London and New York. Routledge

Rothbauer P (2008) Triangulation. The SAGE encyclopedia of qualitative research methods 1(12):892–894

Sag IA (2012) Sign-based construction grammar: an informal synopsis. In: Boas HC, Sag, IA (eds) Sign-based construction grammar. CSLI publications, Stanford, pp 69–189

Salvucci G, Salvati L (2022) Official statistics, building censuses, and OpenStreetMap completeness in Italy. Int J Geo Inf 11(1):1–29

Samo G, Ursini F-A (2022) Exploring dynamic on-line gazetteers to map variation in the syntax of Italian urbanonyms. Quaderni di lavoro ASIt 24(1):407–423

Samo G, Ursini F-A (2024) Dictionnaire et atlas : propriétés lexicales et sémantiques des urbanonymes en français. Onoma 59(3):277–303

Samo G, F-A Ursini F-A (2023) Geographical maps meet place names where languages meets dialects: the case of Italian. Forum Ital 57(3)

Sarkar D, Anderson JT (2022) Corporate editors in OpenStreetMap: investigating co-editing patterns. Trans in GIS 26(4):1879–1897

Schäfer T, Kieslinger B (2016) Supporting emerging forms of citizen science: a plea for diversity, creativity and social innovation. J Sci Com 15(02):Y02

Senaratne H et al. (2017) A review of volunteered geographic information quality assessment methods. Int J Geogr Inform Sci 31(1):139–167

Seto T (2022) Development of OpenStreetMap Data in Japan. In: VV.AA (ed). Ubiquitous mapping. Springer, Berlin, pp. 113–126

Southall H, Aucott P (2019) Expressing history through a geo-spatial ontology. Int J Geo Inf 8(2):362–379

Stolz T, Ingo H, Warnke IH (2018) Comparative colonial toponomastics: evidence from German and Dutch colonial placenames. In: Sonstige Namenarten. Stiefkinder der Onomastik. de Gruyter, Berlin & Boston, pp 90–108

Touya G et al. (2017) Assessing crowdsourced POI quality: combining methods based on reference data, history, and spatial relations. Int J Geo-Inf 6(3):60–80

Tuan Y-F (1977) Space and place: the perspective of experience. University of Minnesota Press

Ursini F-A, Zhang YS (2023) Place and place names: a unified model. Front Psychol 14:1237422

Ursini F-A (2020) National identities and collective memory in Italian toponyms. In: Nick IA, Piccozzi M (eds.), Re)naming places, (Re)shaping identities. Cambridge Scholars Publishing, Cambridge, pp 173–186

Ursini F-A, Samo G (2023) Extracting toponyms from OpenStreetMap: a cross-linguistic perspective. In: Xuke H, Hu Y, Kerstens J, Stock K (eds) Proceedings of GeoExt2023. Springer, Berlin 110–120

Vannieuwenhuyze B (2007) The study and classification of medieval urban toponymy: the case of late medieval Brussels (13th-16th centuries). Ono 42(1):189–211

Walkowiak J (2024) City-text as mapping a territory: 'Polish' streets in Berlin. Natl Pap 1–19

Way T (2019) What is the urban landscape and what role in urban history? J Urb His 45(3):595–600

Westerholt R (2019) Methodological aspects of the spatial analysis of geosocial media feeds: From locations towards places. GIS Sci 31(1):65–76

Westerholt R (2019) The analysis of spatially superimposed and heterogeneous random variables. PhD Dissertation, University of Heidelberg,

Wu K, Xie Z, Hu M (2022) An unsupervised framework for extracting multilane roads from OpenStreetMap. Int j Geo Inf Sci 36(11):2322–2344

Xie S, Ursini F-A, Samo G (2023) Urbanonyms in Macau. Names 71(1):29–43

Yee H (2014) The theory and practice of one country, two systems in Macau. In: Yu EWY, Chan MK China's Macao transformed: challenge and development in the 21st century. City University of Hong Kong Press, Hong Kong, pp 3–20

Zhou Q et al. (2022) Assessing OSM building completeness for almost 13,000 cities globally. Int J Dig Ear 15(1):2400–2421

Author contributions

F-AU has designed the manuscript, analysed the data and prepared the first draft. GS has collected and analysed the data, and prepared the supplementary materials and figures.

Both authors have equally contributed to the preparation and final versions of the manuscript.

Competing interests

The authors declare no competing interests.

Ethical approval

This article does not contain any studies with human participants performed by any of the authors.

Informed consent

This article does not contain any studies with human participants performed by any of the authors.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1057/s41599-025-05025-1.

Correspondence and requests for materials should be addressed to Francesco-Alessio Ursini or Giuseppe Samo.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.