The Semantics of Generic Terms in Toponyms: Formal Semantics Meets Multi-source Big Data

Giuseppe Samo[1,†], Francesco-Alessio Ursini[2,*,†]

[1]Beijing Language and Culture University, Xue Yuan Road 15, 100083 Beijing, China

[2]Central China Normal University, School of Chinese Language and Literature, 52, Dailyou Road, Wuhan, 625762, China

Abstract

The goal of this paper is to offer a formal semantic analysis of generic terms in toponyms (e.g. *square* in *Trafalgar Square*). The analysis builds on multi-source big data. Data are collected from multiple textual sources (corpora, LLMs, ConversationalAI based on LLMs, and dictionaries): the size aims to cover a statistically significant portion. One Italian generic term, *piazza* 'square', is used as a benchmark for the analysis due to its widespread usage and consequent semantic flexibility. The formal analysis integrates a variant of first order predicate logic, Discourse Representation Theory, with a frame semantics-based analysis distinguishing context-sensitive uses of words (i.e. general, specialised, technical term uses). It shows that different uses of *piazza* may correspond to semantically related structured formulae. Their content can thus be compared via the definition of a semantic overlap function. The paper concludes by discussing consequences for a semantic theory of generic terms and toponyms.

Keywords

Toponyms, Generic Terms, multi-source data, Geosemantics, Italian,

1. Introduction

Geographic Information Science (GIS) has studied the conceptual and linguistic properties of place names or, alternatively, *toponyms* from a variety of perspectives. Several studies analyse toponyms' grammatical properties and their internal composition [1, 2], their identification/retrieval in texts [3, 4], and disambiguation in context [5, 6]. Most studies, however, study toponyms' potential content, and mostly follow two strands of research: the psychological/conceptual and the semantic strands. Both strands are centred on the notion of *place*: geographical entities that agents can imbue with emotional, cognitive and social content [7, 8]. These strands take different but related approaches to the constituting elements of this content, and to which aspects of cognition they belong. These theoretical differences can be summarised as follows.

Psychologically oriented works focus on how toponyms can convey non-linguistic and subjective concepts about specific places (e.g. a city such as Rome) and kinds of places (e.g. cities) [9, 10]. They often propose that places correspond to complex mental/cognitive representations,

GeoExT 2024: Second International Workshop on Geographic Information Extraction from Texts at ECIR 2024, March 24, Glasgow, Scotland

[*] Corresponding author.

[†] These authors contributed equally.

samo@blcu.edu.cn (G. Samo); randorama@outlook.com (F. Ursini)
https://github.com/samo-g (G. Samo)
https://orcid.org/0000-0003-3449-8006 (G. Samo); 0000-0001-7042-3576 (F. Ursini)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

only partially amenable to formalisation (cf. the ample discussion in [11]). Instead, semantically oriented works focus on the linguistic content of toponyms [12, 13], e.g. under which conditions toponyms and their constituents (e.g. *square* in *Trafalgar Square*) can describe place types [14]. Such formalizations are implementable in formal ontologies, usually via variants of first order logic (e.g. DOLCE, formal concept analysis [15]). Instead, psychological approaches permit researchers to model place concepts in a flexible manner via the use of *facets*, mental features representing mind-external but agent-oriented aspects of places [16, 17].

Computationally-driven GIS models have integrated these strands via two key steps [18, 19]. First, one defines the facets identifying places by extracting place descriptions from texts (e.g. the facet 'beautiful' for London, from a TripAdvisor review). Second, one adopts an ontological commitment: place concepts represent objects that occupy locations in which events can occur, and they are defined via facets. The linguistically oriented [20] offers a formalisation of place concepts so defined via *Discourse Representation Theory*, a variant of first order logic (DRT: [21, 22]). Place concepts combine dynamic variables (or *referents*) belonging to various semantic types, and conditions/facets individuating these referents. The proposal suggests that toponyms are unique linguistic labels for place concepts, and thus integrates psychological, ontological and semantic models of place concepts into one framework (cf. also [23]).

The recent [24] shows that content and context of use for toponyms systematically correlate. The proposal follows terminology studies and suggests that words can have at least three context-driven uses (cf. also [25, 26]). In the *general use*, agents use words in a non-qualified manner (e.g. *cat* in daily chats). In the *specialised* use, agents use words often in situation-specific manners (e.g. *cat* in a zoology lecture). In the *(technical) term* use, agents associate words' content with necessary and sufficient conditions of use, based on a given set of texts (e.g. *cat* used in pedigree evaluation documents). The proposal suggests that generic terms in toponyms (e.g. *alley*) can have each of these three uses, offering corpus- and dictionary-based evidence in support. The proposal formalises these uses via a frame semantics approach ([27, 28]), and shows that uses share some content: they stand in a semantic overlap relation.

Our goal in this paper is to integrate these proposals into a unified model. Empirically, we analyse different textual sources of increasingly restricted use: corpora, LLMs and ConversationalAI based on LLMs for general use; wikipedia articles for specialised uses; technical dictionaries for term uses. We focus on the Italian generic term *piazza* 'square' because Italian offers a wealth of textual sources for data extraction, and Italian toponyms are well-studied ([29, 30]), but lack formal semantic analyses.[1] We derive facets' sets corresponding for each use in the selected sources. Theoretically, we test the unified model by showing how the uses of *piazza* can overlap in their content. We conclude by discussing consequences for a theory of toponyms' conceptual/semantic content.

2. Data retrieval

Our methodology is summarized in flowchart 1 and explains how the different sources determine the data for the analysis. The type of data we analyse are naturally occurring sentences in

[1] Toponyms can include various types of generic terms, some of them not having spatial/platial content (e.g. *new* in *New South Wales*). We leave an analysis of these cases for future research.

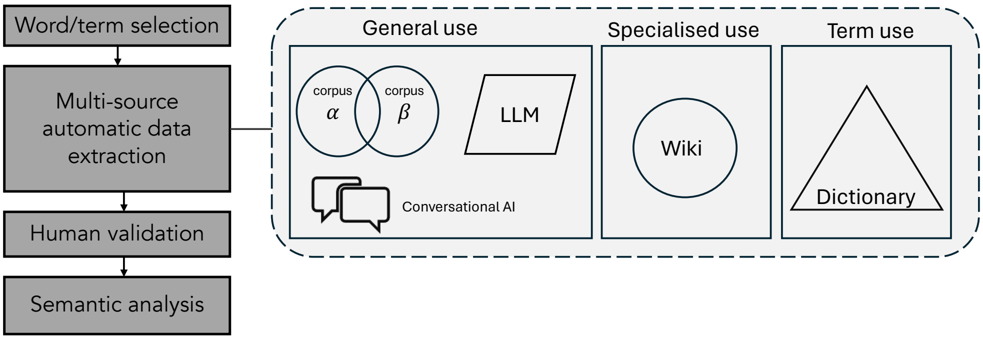

Figure 1: Flowchart: After selecting words/terms, we retrieved data from different sources in an automated manner. We then proceeded with the human validation and semantic analysis steps. The details of the data extraction process and the semantic analysis are provided in the running text.

corpora, Large-Language Models (LLMs), conversational AI, wikipedia articles, and dictionary entries. From these sentences, we extract toponyms including *piazza* and analyse their content via co-occurring adjectives and nouns (e.g. *grande* 'big', *duomo* 'cathedral'). Before we offer the results, we present the formal properties of each source.[2]

Corpora: We first explore how the term *piazza* is used in corpora, as they offer naturally occurring examples for general uses. Building on [24], we rely on large-scale corpora: we have decided to cross two types of web-based datasets for Italian, the corpus Paisà ([31], ca. 250 millions tokens) and news, as the corpus Repubblica ([32], ca. 325 millions tokens). Both corpora allow good accessibility for data extraction. We queried both platforms[3] and downloaded the queried results as .txt files listing sentences including *piazza* (respectively, 24,076 sentences for Paisà; 77,586 Repubblica). We then ran a python script first introduced in [24] to automatically find collocations. We thus explored the metrics of the Pointwise Mutual Information [33] that maps semantic relations between target words and words collocating with them [34, 35].

We removed common words in lower cases to remove noise (e.g. *piazza* used as a form of the verb *piazzare* 'to place'), false positive tokens (e.g. the family name *Cavour*, which often occurs in the toponym *Piazza Cavour* and commemorates the first Prime Minister of Italy), and unrelated uses due to polysemy (e.g. *piazza affari* 'business square', the name of Milan's stock market and not a literal square). Following [24], for the specialized uses we explored wikipedia pages focusing on squares and thus including the target word *piazza*. We ran a scraped content from 1000 pages with at least one occurrence of *piazza* in the text and created a .txt file, which is then analyzed similarly to the previous datasets.

LLM and Conversational-AI based on LLM: Large Language Models (LLMs) can be used in the automatic retrieval of semantic content of toponyms. Recent studies indicate that language models (e.g. Neural Networks and LLMs, see [36] for a definition) demonstrate proficiency in various linguistic tasks [37], although we have mixed findings regarding their encoded semantic knowledge [38]. We investigate transformer-based language models through fill-masking, a

[2] All materials (scripts, outputs) are available at the following link: https://github.com/samo-g/piazza

[3] Paisà: https://www.corpusitaliano.it/ (last access 17.01.2024);
Repubblica: https://bellatrix.sslmit.unibo.it/noske/public/#dashboard?corpname=repubblica (last access 17.01.2024)

task involving masking an n-gram (word) in a devised ex-novo sentence and prompting the model to predict suitable replacements for the masked region. The prompting aims to retrieve specific part-of-speech elements (i.e. noun, adjective, verbs): for instance, *questa piazza è molto [MASK]* 'this square is very [MASK]' is one of the inputs to retrieve adjectives [39].

The model outputs a probability score, where higher values signify a greater likelihood of a given n-gram being appropriate in a particular sentence. We interpret higher probabilities as indicative of stronger semantic associations, suggesting that a word is more likely to convey the intended meaning in the context of a sentence. Specifically, we employed the transformer-based language model DBMDZ/BERT-BASE-ITALIAN-UNCASED.[4] The training set also contains wikipedia dump, but the heterogeneous nature of the corpus lead us to consider the results are equivalent to general use. We analyzed the output consisting of 120 datapoints and their score.

Finally, in the spirit of [40], we aimed to interact with Chat-GPT 3.5 (last access 17.01.2024; [41], [42] *inter alia*; see an overview and critical discussions in [43]) in a series of "functional" interactions, i.e. extremely general questions with functional elements as bare interrogatives and auxiliaries, e.g. *Cosa è una piazza?* 'what's a square?' or *cosa posso fare in una piazza in Italia?* 'what can I do in a square, in Italy?'. All interactions were in standard Italian. We considered all the interactions with Chat-GPT as general use. We analysed the outputs of the conversational AI as a corpus [44], for a total of 1,489 tokens.

Dictionaries: Supported by previous cross-linguistic literature [45], we retrieved the definition offered by the dictionary Treccani[5]. We chose this typology of definition, since it sheds light on the the social functions of the term *piazza*.

In table 1 we list all the retrieved items: we assume that they describe facets of content associated with *square* and contextualised to a given source/use (corpora for general uses).

3. Analysis

We assume that toponyms are a sub-type of proper names, and thus of phrasal coordinated compounds [46, 47]. Italian toponyms including *piazza* as a generic term are thus the combination of a generic and a specific term (i.e. *Piazza Cavour* includes generic *Piazza* and specific *Cavour*). The semantic content of toponyms corresponds to the combined content of generic and specific terms, plus the naming/etymological relation connecting these terms. In the previous section, we have shown that generic terms minimally contribute lists/sets of features that can describe place types (here, squares). Specific terms introduce reference to entities, locations or historical/cultural events that are commemorated via a given toponym (e.g. *Cavour* being the surname of the Italian prime minister to which the square is dedicated).

In order to pursue a unitary representation of grammatical form and content, [20] uses DRT as formal framework. In DRT, the content of linguistic units (from morphemes to discourses) is represented via *Discourse Representation Structures* (DRSs), a variant of formulae of first order logic [21, 22]. DRSs pair a set of referents (the *universe of discourse*) with conditions (properties, relations) individuating these references. The dynamic status of referents hinges on their use in the building of DRSs, as we show via the DRSs in Fig. 2.

[4] https://huggingface.co/dbmdz/bert-base-italian-uncased (last access, 24.01.2024; 13GB and 2,050,057,573 tokens).
[5] (https://www.treccani.it/enciclopedia/piazza_%28Enciclopedia-Italiana%29/, last access 20.01.2024.

Source	**word and gloss/ facet**
GENERAL	
Corpora	*ampia* ’wide’, *antica* ’ancient’, *bella/bellissima* ’beautiful/very beautiful’, *cattedrale* ’cathedral’, *castello* ’castle’, *celebre* ’celebrated’, *centralissima* ’very central’, *cinquecentesca* ’from the 1500s’, *commerciale* ’commercial’, *conosciuta* ’known’, *deserta* ’deserted’, *ducale* ’ducal’, *duomo* ’cathedral’, *elegante* ’elegant’, *eroi* ’heroes’, *famosa* ’famous’, *grande* ’big’, *gremita* ’crowded’, *importante* ’important’, *invasa* ’invaded’, *italiana* ’Italian’, *luminosa* ’bright’, *magnifica* ’magnificent’, *medievale/medioevale* ’medieval’, *mercato* ’market’, *monumentale* ’monumental’, *municipio* ’city hall’, *museo* ’museum’, *nostra* ’our’, *palazzo* ’palace’, *pedonale* ’pedestrian’, *percorrendo* ’traversing’, *principale* ’main’, *rinascimentale* ’Renaissance’, *risorgimento* ’resurgence’, *stazione* ’station’, *tranquilla* ’quiet’, *vuota* ’empty’
LLM	*hotel* ’hotel’, *chiesa* ’church’, *statua* ’statue’, *bella* ’beautiful’, *tranquilla* ’quiet’, *grande* ’big’, *interessante* ’interesting’, *centrale* ’central’, *attraente* ’attractive’, *curata* ’well-kept’, *mangiare* ’eating’, *passeggiare* ’walking’, *giocare* ’playing’, *visitare* ’visiting’
Chat-GPT 3.5	*aperta* ’open’, *spaziosa* ’spacious’, *incontro* ’meeting’, *commerciale* ’commercial’, *fontane* ’fountains’, *monumenti* ’monuments’, *vita sociale* ’social life’, *vita culturale* ’cultural life’, *accessibilità* ’accessibility’, *spaziosa* ’spacious’, *mercato* ’market’, *arte* ’art’, *gelato* ’ice cream’, *caffè* ’coffee’
SPECIALIZED	
Wikipedia	*architettura* ’architecture’, *fontana* ’fountain’, *pubblica* ’public’, *principale* ’main’, *alberata* ’tree-lined’, *importante* ’important’, *organizzare* ’to organize’, *quadrangolare* ’quadrangular’, *municipio* ’city hall’
TERM USE	
Treccani	*spazio libero* ’open space’, *delimitato da costruzioni* ’delimited by buildings’, *funzione politica* ’political function’, *funzione religiosa* ’religious function’, *funzione commerciale* ’commercial function’

Table 1
Italian words (and their English glosses) representing facets, sources and use.

A man walks in a park. He whistles.
(Adapted from [21], (16))

x y z t
man’(x)
walk’(x)
in’(x,t)
t:**loc**’(y)
park’(y)
whistle’(z)
man’(z)
z = x

Sydney

s^{o} $\mathrm{p}^{\lambda,\, n\, =\, 1}$ S^{ϕ}	
p:**loc**’(s)	
S’(p) =	$\sqcup$ {**extension**’(p)
	number-of-districts’(p)
	popularity’(p)
	citizens-opinion’(p)}

Figure 2: Examples of DRSs.

The two-sentence discourse in Fig. 2, left side, establishes that there is a man walking in a park (first sentence) and whistling (second sentence). The referent x represents a walking man, the referent t represents the park’s location, and the referent y represents the park as a physical entity occupying this location. The referent *z* represents a whistling man that, via the anaphoric nature of pronoun *he*, co-refers to the walking man introduced in the first sentence. This identity is dynamically established via the relation *z*=*x*, which binds two distinct referents as representing the same entity. The condition ***loc’*** dynamically binds the referent *y* for the park as a physical entity with the location *t* that this park occupies (i.e. we have *t*:***loc’****(y)*). Conditions describing the physical entity (i.e. ***park’***) also describe the corresponding location. Referents have default existential import: for instance, *x* and *y* in the universe of discourse stand short for $\exists x$, $\exists y$. The mental content associated to a discourse can thus be represented as a complex DRS, by composing together the concepts/DRSs that each word in a discourse contributes.

The content associated to a toponym (here, *Sydney*) can be represented via the DRS in Fig. 2, right side. The referent *s* represents Sydney as a complex yet bounded object (e.g. an agglomeration of buildings). The condition **loc’** maps this object onto a place *p*: a *unique location* for an object (viz. the superscript $n = 1$ for the *p* referent). The content associated to this place referent is the (Boolean) join set of conditions in the DRS (e.g. Sydney’s extension as ***extension’****(p)*), represented via the symbol ⊔ (roughly, set union on conditions). Thus, the semantic content of *Sydney* corresponds to a complex description of the conditions (i.e. facets) identifying this city (i.e. we have ***S’****(p)*=⊔***Facet’***$_n$*(p)*). Referents are assigned to types, i.e. ontological categories: *s* belongs to the type o of objects, *p* to the type λ of locations, and *S* to the type ϕ of facets (conditions, in DRT). The proposal thus analyses toponyms’ content via DRSs: conditions represent facets, and referent types represent categories of the ontological commitment in [19, 20].

The proposal sketches an approach to toponyms’ use in context as being potentially flexible; however, this implementation remains underdeveloped. The proposal in [24] focuses on this latter aspect, as we show in Table 2 via the frames for *square*:

	General use	Specialized use	Term use
	[*parts:signs*]⊔	[*function:market*]⊔	[*place_type:intersection_of_streets*] ⊓
	[*use:pedestrian*]⊔	[*use:pedestrian*]⊔	[*use:pedestrian*]⊓
	[*access:open*]⊔	[*function:commercial*]⊔	[*parts:grass*]⊓
	[*service:parking*]⊔	[*type:monumental*]⊔	[*parts:trees*]⊓
ATTRIBUTES =	[*parts:lamps*]⊔	[*location:central*]⊔	[*function:park*]⊓
	[*users:artists*]⊔		[*location:next_to_buildings*]⊓
	[*etym:commemoration*]⊔	[*etym:commemoration*]⊔	[*etym:commemoration*]⊓
	[*commem_entity:battle*]⊔	[*commem_entity:battle*]⊔	[*commem_entity:battle*]⊓
	[*C*(ontext attributes)]⊔	[*C*(ontext attributes)]⊔	[*C*(ontext attributes):
	...	...	x (e.g., legal context)] ⊓ ...

Table 2
Frame for *square*, from [24], page 7. The results are restricted to the content extracted for the toponym *Trafalgar Square*, for ease of illustration.

The content of toponyms is modelled as attribute-value matrices that represent words’ content (i.e. the facets describing places that they convey), relativized to their context of use (cf. also [48, 49, 50, 51, 52]). For instance, [*use:pedestrian*] and [*function:market*] in the specialised vocabulary use (central frame, Table 1) are pairs representing that squares have pedestrian uses and market functions (i.e. we treat features as values of general attributes). An attribute representing context- and case-sensitive uses, *C*, can be used as an open slot for emergent content in context [53]. The frames in the table are referent-free structures: they do not specify which referents and corresponding relations form places so defined.

General and specialised uses differ from term uses regarding the properties of their sets of pairs/facets. General and specialised uses involve join sets of pairs (viz. the symbol $\sqcup$); term uses involve meet sets of pairs represented via the symbol $\sqcap$. Lists of pairs vary, based on the facets individuated in context. Furthermore, pairs forming general uses come from distinct sentences. In one sentence one can use *square* for a pedestrian place; in another, for a place used by artists. Specialised uses usually involve longer texts attributing several facets at once to a place, though these facets are also interpreted disjunctively. Via meet, frames are interpreted conjunctively: a square is described as having all listed properties necessarily at once.

The unification of the two approaches can be achieved via three steps. We show how this is the case via the formalisation of *piazza* in Table 3. We also show how this formalisation extends to toponyms by offering data based on the semantic analysis of *Piazza San Marco*, the name of the famous square in Venice:

	General use	Specialized use	Term use
	x^{o} p^{λ} P^{ϕ}	x^{o} p^{λ} P^{ϕ}	x^{o} p^{λ} P^{ϕ}
Facet-set=P=	$\{$*p:**loc’**(x)* $\sqcup$	$\{$*p:**loc’**(x)* $\sqcup$	$\{$*p:**loc’**(x)* $\sqcap$
	architectural$_{\textbf{part}}(p)]\sqcup$	**gathering**$_{\textbf{function}}(p)\sqcup$	**buildings**$_{\textbf{part}}(p)\sqcap$
	pedestrian$_{\textbf{use}}(p)\sqcup$	**pedestrian**$_{\textbf{use}}(p)$	**open**$_{\textbf{access}}(p)\sqcap$
	open$_{\textbf{access}}(p)\sqcup$	**commercial**$_{\textbf{function}}(p)\sqcup$	**political**$_{\textbf{function}}(p)\sqcap$
	leisure$_{\textbf{service}}(p)\sqcup$	**monumental**$_{\textbf{type}}(p)\sqcup$	**commercial**$_{\textbf{function}}(p)\sqcap$
	cons$_{\textbf{part}}(p)\sqcup$	**central**$_{\textbf{location}}(p)\sqcup$	**religious**$_{\textbf{function}}(p)\sqcap$
	tourists$_{\textbf{users}}(p)\sqcup$		**next_to_buildings**$_{\textbf{location}}(p)\sqcap$
	commemoration$_{\textbf{etymology}}(p)\sqcup$	**commem**$_{\textbf{etymology}}(p)\sqcup$	**commem**$_{\textbf{etymology}}(p)\sqcap$
	saint$_{\textbf{commem\_entity}}(p)$ $\sqcup$	**saint**$_{\textbf{commem\_entity}}(p)\sqcup$	**saint**$_{\textbf{commem\_entity}}(p)\sqcap$
	C(ontext conditions)*(p)*$\}$	*C*(ontext conditions)*(p)*$\}$	*C*(ontext conditions)*(p)* $\}$

Table 3
DRSs/frames for the uses of*Piazza San Marco*.

Frames and DRSs become near-equivalent structures via two steps. The first step is the introduction of referents and their dynamic relations. Each DRS includes the condition *p:**loc’**(x)* as their initial condition/attribute: generic words/terms implicitly refer to places in any context of use. Therefore, each condition/attribute representing a facet individuates a property of a place as a unique object in a location. The second step is the introduction of types for conditions. Conditions represent attribute-value pairs via the format ***value**$_{\textbf{attribute}}$*(v): a condition corresponds to a value for an attribute as a semantic type, assigned to a referent *v*. For instance, the condition ***gathering**$_{\textbf{function}}$*(p) represents the fact that squares can be places where individuals freely gather for various purposes. Conditions so defined thus also represent the fact that more abstract facets (as attributes) can have more concrete realisations (i.e. values).

Uses and corresponding DRSs also differ in subtle manners. The left and central DRSs differ in the conditions they include, but also in how these conditions are computed. Each condition in the left DRS comes from one sentence providing a token for *piazza*: again, general uses tend to focus to one or few facets describing a place. Several of the conditions in the central DRS however co-occur in one or more texts, e.g. that squares can have gathering and commercial uses. Specialised uses are such also because they are defined over larger, cohesive bodies of text(s). *Piazza* as a technical, generic term corresponds to the right DRS, which includes the meet of the various facets/conditions describing squares. We can thus represent the content of generic terms and their contribution of place names (in this case, *Piazza San Marco*) via DRT plus conditions on context of use. Again, conditions/facets may also emerge in the contexts of use of this toponym: in each DRS, The condition *C(ontext conditions)* represents this possibility.

As [24] also shows, uses can be compared via an overlap semantic relation, $\circ$. This relation holds when two formal structures A and B have at least one shared element (here, condition sets, i.e. $A \sqcap B \neq \emptyset$), but their union/join forms a different set (i.e. $A \sqcup B = C$). In our case, the set of conditions $P' = \textit{piazza\_gen} \circ \textit{piazza\_spec} \circ \textit{piazza\_term} = \{p{:}\textit{\textbf{loc'}}(x), \textit{\textbf{commem}}_{\textit{\textbf{etymology}}}(p), \textit{\textbf{saint}}_{\textit{\textbf{commem\_entity}}}(p)\}$ defines those facets that agents ascribe to Italian *piazza San Marco* irrespective of the context. The three uses minimally overlap in at least three facets, respectively: it is a place as a complex referent, it bears a name commemorating another referent, and this referent is a saint. Specialised and term uses also overlap in describing this square as having commercial functions (i.e. we have the set $P'' = \textit{piazza}_{spec} \circ \textit{piazza}_{term} = \{p{:}\textit{\textbf{loc'}}(x), \textit{\textbf{commercial}}_{\textit{\textbf{function}}}(p), \textit{\textbf{commem}}_{\textit{\textbf{etymology}}}(p), \textit{\textbf{saint}}_{\textit{\textbf{commem\_entity}}}(p)\}$). More in general, uses can overlap in various manners, depending on which facets emerge via these uses. Our unified model can represent this fact, and thus some agents can use toponyms to describe the same place in different contexts, and that these uses share some content.

4. Discussion

We believe that two overarching results emerge from our proposal. First, we present an automated methodology for collecting geolinguistic data from large-scale repositories and LLMs (cf. [34, 35]). Beyond standard frequency and collocation analyses, we have developed a method to explore masked modeling and interactions with conversational AI (cf. [36, 37]). Despite the challenges in grasping the generalizability of synthetic/simulation-based data retrieved from LLMs and conversational AI with respect to human results (see [54], but also [55]), the evident asymmetries underscore their significance as valuable linguistic data. The methodology also hinges on a multi-source collection of data, thus achieving a broader and more balanced empirical base (cf. [56]). Future studies should focus on developing prompt settings for retrieving specialized and technical term uses, if feasible.

Second, we present a model that integrates DRT and frame semantics approaches to toponyms (cf. [20, 24]). This model combines insights from formal semantic models (e.g. [12, 13]), ontologically committed models (e.g. [18, 19]) and psychologically oriented models (e.g. [10, 16, 17]). Furthermore, this integration can capture different contexts of use (i.e. general, specialised, term: [25, 26]) by representing their differences in interpretation and content associated to each use. The model thus achieves several unification goals via a relatively minimal theoretical apparatus. Future studies should develop further aspects of the theoretical apparatus outlined in this paper (e.g. places and their ontologies: [57, 58]).

In conclusion, this paper has offered a unified model for the content of toponyms that combines a DRT account with a frame semantics-based analysis of this content. The resulting structures, or DRSs, are indexed with respect to three contexts of use (general, specialised, term), and compared via the definition of a semantic overlap relation. The empirical reach of the model builds on evidence collected from multiple sources (e.g. corpora, wikipedia articles, dictionaries) that define these three contexts of use. The unification of "big data" data analysis with an integrated formal model passes via the analysis of the Italian generic word/term *piazza* 'square', and thus can potentially apply to the tens of thousands of toponyms including this term in this language. The model thus informs research in GIS, but also in toponomastics and psychology of place, and may possibly be extended in future work.

References

[1] F. Perono Cacciafoco, F. Cavallaro, Place Names: Approaches and perspectives in toponymy and toponomastics, Cambridge: Cambridge University Press, 2023.

[2] W. Kuhn, Core concepts of spatial information for transdisciplinary research, International Journal of Geographical Information Science 26(12) (2012) 2267–2276. doi:https://doi.org/10.1080/13658816.2012.722637.

[3] M. Vasardani, S. Winter, K.-F. Richter, Locating place names from place descriptions, International Journal of Geographical Information Science 27(12) (2013) 2509–2532. doi:https://doi.org/10.1080/13658816.2013.785550.

[4] H. Chen, M. Vasardani, S. Winter, Georeferencing places from collective human descriptions using place graphs, Journal of Spatial Information Science 17(1) (2018) 31–62. doi:http://dx.doi.org/10.5311/JOSIS.2018.17.417.

[5] D. Buscaldi, Toponym Disambiguation in Natural Language Processing, Ph.D. thesis, University of Valencia, Valencia, 2011.

[6] C. Davis, Reading geography between the lines: Extracting local place knowledge from text, in: A. G. . Z. W. E. Thora Tenbrink, John Stell (Ed.), International Conference on Spatial Information Theor, Springer, Berlin, 2013, pp. 320–337. doi:https://doi.org/10.1007/978-3-319-01790-7_18.

[7] T. Cresswell, Place: an introduction, John Wiley & Sons, 2014.

[8] J. Malpas, Place and Experience: A philosophical topography, Routledge, Sidney, 2018.

[9] J. S. Smith (Ed.), Place and Experience: A philosophical topography, Routledge, London, 2017. doi:doi:10.4324/9781315189611.

[10] L. C. Manzo, P. Devine-Wright (Eds.), Place Attachment: Advances in theory, methods and applications, Routledge, New York, 2021.

[11] F.-B. Mocnik, Putting Geographical Information Science in Place – Towards Theories of Platial Information and Platial Information Systems, Progress in Human Geography 46 (2022) 798–828. doi:10.1177/030913252210740.

[12] P. Yue, Geospatial semantic web, in: P. Yue (Ed.), Semantic web-based intelligent geospatial web services, Springer, New York, 2013, pp. 17–20. doi:https://doi.org/10.1007/978-1-4614-6809-7_3.

[13] Y. Hu, Geospatial semantics, in: B. Huang, T. J. Cova, M.-H. T. et al. (Eds.), Comprehensive Geographic Information Systems vol. 1, Elsevier, Oxford, 2018, pp. 80–94. doi:http://dx.doi.org/10.1016/B978-0-12-409548-9.09597-X.

[14] K. Kijania-Placek, Names of Places, Semiotica 240(1) (2018) 187–210. doi:https://doi.org/10.1515/sem-2021-0020.

[15] A. Ballatore, Prolegomena for an ontology of place, in: H. Onsrud, W. Kuhn (Eds.), Advancing Geographic Information Science, GSDI Association Press, Oxford, 2016, p. 91–103.

[16] D. Canter, The facets of place, in: G. T. Moore, R. W. Marans (Eds.), Toward the integration of theory, methods, research, and utilization, Springer US, Boston, MA, 1997, p. 109–147. doi:https://doi.org/10.1007/978-1-4757-4425-5_4.

[17] D. Canter, Facet theory: Approaches to social research, Springer Science Business Media, 2012.

[18] S. W. Purves, Ross, W. Kuhn, Places in information science, Journal of the Association for Information Science and Technology 70(11) (2019) 1173–1182. doi:doi:10.1002/asi.24194.

[19] S. W. Hamzei, Ehsan, M. Tomko., Place facets: A systematic literature review, Spatial Cognition Computation 20(1) (2020) 33–81. doi:doi:1080/13875868.2019.1688332.

[20] F.-A. Ursini, Y. Zhang, Place and place names: a unified model, Frontiers in Psychology 14 (2023) nn. doi:doi:10.3389/fpsyg.2023.1237422.

[21] J. v. G. Kamp, Hans, U. Reyle, Discourse representation theory, in: D. Gabbay, F. Gunthner (Eds.), Handbook of Philosophical Logic 15, Kluwer, Dordrecht, 2011, p. 125–394.

[22] H. Kamp, The links of causal chains, Theoria 88(2) (2022) 296–325. doi:https://doi.org/10.1111/theo.12381.

[23] T. Tenbrink, The language of place: towards an agenda for linguistic platial cognition research, ?, Heidelberg, 2020, pp. 5–12. doi:10.5281/zenodo.3628849.

[24] F.-A. Ursini, G. Samo, A semantic model for generic terms and place nouns, in: AA.VV. (Ed.), Proceedings of SDSS 2023, 2023, pp. 1–10.

[25] P. ten Hacken, Terms and specialized vocabulary: Taming the prototypes, John Benjamins, Amsterdam & Philadelphia, 2015, pp. 3–13.

[26] P. ten Hacken, Terms between standardization and the lexicon, Roczniki Humanisticyzne 66 (2018) 100–118. doi:https://doi.org/10.18290/rh.2018.66.11-4.

[27] R. Naumann, An outline of a dynamic theory of frames, Springer, Berlin, 2013, pp. 26–30.

[28] T. Gamerschlag, D. Gerland, R. Osswald, W. Petersen (Eds.), Meaning, frames, and conceptual representation, Düsseldorf University Press, Düsseldorf, 2015.

[29] G. Samo, F.-A. Ursini, Exploring dynamic on-line gazetteers to map variation in the syntax of italian urbanonyms, Quaderni di lavoro ASIt 24 (2023) 407–423.

[30] G. Samo, F.-A. Ursini, Geographical maps meet place names where languages meets dialects: The case of italian., Forum Italicum 57(3) (2023) 1019–1040. doi:10.1177/00145858231190030.

[31] C. Borghetti, C. S, B. M, I testi del web: una proposta di classificazione sulla base del corpus paisÀ, in: E. . O. C. Cerruti, M. / Corino (Ed.), Scritto e parlato, formale e informale: La comunicazione mediata dalla rete., Carocci, Rome, 2011, pp. 147–170.

[32] M. Baroni, S. Bernardini, F. Comastri, L. Piccioni, A. Volpi, G. Aston, M. Mazzoleni, Introducing the la repubblica corpus: A large, annotated, TEI(XML)-compliant corpus of newspaper Italian, in: M. T. Lino, M. F. Xavier, F. Ferreira, R. Costa, R. Silva (Eds.), Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04), European Language Resources Association (ELRA), Lisbon, Portugal, 2004. URL: http://www.lrec-conf.org/proceedings/lrec2004/pdf/247.pdf.

[33] K. Church, P. Hanks, Word association norms, mutual information, and lexicography, Computational linguistics 16 (1990) 22–29.

[34] J. A. Bullinaria, J. P. Levy, Extracting semantic representations from word co-occurrence statistics: A computational study, Behavior research methods 39 (2007) 510–526.

[35] G. Recchia, M. N. Jones, More data trumps smarter algorithms: Comparing pointwise mutual information with latent semantic analysis, Behavior research methods 41 (2009) 647–656.

[36] J. Hale, Information-theoretical complexity metrics, Language and Linguistics Compass 10 (2016) 397–412.

[37] T. Linzen, M. Baroni, Syntactic structure from deep learning, Annual Review of Linguistics 7 (2021) 195–212.

[38] E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, On the dangers of stochastic parrots: Can language models be too big?, in: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, 2021, pp. 610–623.

[39] G. Samo, F.-A. Ursini, Large Language Models, Frame Semantics and Geodialectal Data, Proceedings of SIDG (to appear).

[40] A. Massaro, G. Samo, Prompting metalinguistic awareness in large language models: Chatgpt and bias effects on the grammar of italian and italian varieties, Verbum 14 (2023).

[41] OpenAI, Gpt-4 technical report. (2023).

[42] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al., Improving language understanding by generative pre-training, Available from: https://www.cs.ubc.ca/ amuham01/LING530/papers/radford2018improving.pdf (last access 24.01.2024) (2018).

[43] Y. Bang, S. Cahyawijaya, N. Lee, W. Dai, D. Su, B. Wilie, H. Lovenia, Z. Ji, T. Yu, W. Chung, Q. V. Do, Y. Xu, P. Fung, A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity, 2023. arXiv:2302.04023.

[44] G. Mikros, AI-Writing Detection Using an Ensemble of Transformers and Stylometric Features, Talk held at Qualico 2023, University of Lausanne, 29.06.2023 (2023).

[45] E. Klaus-Jürgen, E. Ballard, D. E. Elswo (Eds.), Encyclopedic Dictionary of Landscape and Urban Planning Multilingual Reference Book in English, Spanish, French, and German, Springer, Berlin, 2010.

[46] B. Schlücker, Von donaustrom zu donauwelle. die entwicklung der eigennamenkomposition von 1600-1900, Zeitschrift für Germanistische Linguistik 48 (2020) 238–268. doi:https://doi.org/10.1515/zgl-2020-2002.

[47] B. Schlücker, T. Ackermann, The morphosyntax of proper names: An overview, Folia linguistica 51 (2017) 309–339. doi:https://doi.org/10.1515/flin-2017-0011.

[48] C. J. Fillmore, Frame semantics, Hanshin Publishing, Seoul, 1982, pp. 111–137.

[49] C. J. Fillmore, J. C. R, M. R. Petruck, Background to framenet, International Journal of Lexicography 16 (2003) 235–250. doi:https://doi.org/10.1093/ijl/16.3.235.

[50] C. J. Fillmore, C. Baker, A frames approach to semantic analysis, Oxford University Press, Oxford, 2010, pp. 313–340. doi:10.1093/oxfordhb/9780199544004.013.0013.

[51] S. Löbner, Evidence for frames from natural language, Springer, Heidelberg, 2014, pp. 23–68. doi:https://doi.org/10.1007/978-3-319-01541-5_2.

[52] S. Löbner, Frames at the interface of language and cognition., Annual Review of Linguistics

7 (2021) 261–284. doi:10.1146/annurev-linguistics-042920-030620.

[53] H. Kamp, B. Partee, Prototype theory and compositionality, Cognition 57 (1995) 129–191.

[54] G. Mikros, A. Koursaris, D. Bilianos, G. Markopoulos, Ai-writing detection using an ensemble of transformers and stylometric features, in: Proceedings of IberLEF 2023, 2023.

[55] S. Stevenson, P. Merlo, Beyond the benchmarks: Toward human-like lexical representations, Frontiers in Artificial Intelligence 5 (2022) 796741.

[56] X. Hu, Z. Zhou, H. Li, Y. Hu, F. Gu, J. Kersten, H. Fan, F. Klan, Location reference recognition from texts: A survey and comparison, arXiv preprint arXiv:2207.01683 (2022).

[57] F.-A. Ursini, G. Samo, Names for urban places and conceptual taxonomies: the view from Italian, Spatial Cognition & Computation 22 (2022) 264–292. doi:10.1080/13875868.2021.1954186.

[58] N. Asher, J. Pustejovsky, A type composition logic for generative lexicon, Springer, Dordrecht, 2013, p. 9–68. doi:https://doi.org/10.1007/978-94-007-5189-7_3.