Studying urban space from textual data: Toward a meth-odological protocol to extract geographic knowledge from real estate ads.

Alicia Blanchi, Giovanni Fusco, Karine Emsellem, Lucie Cadorel

Studying urban space from textual data: Toward a methodological protocol to extract geographic knowledge from real estate ads.

Alicia Blanchi[1], Giovanni Fusco[2], Karine Emsellem[3], Lucie Cadorel[4]

[123] Université Côte d'Azur, CNRS, ESPACE -Nice, France
[4] Université Côte d'Azur, INRIA, CNRS, I3S – Sophia-Antipolis, France
alicia.blanchi@etu.univ-cotedazur.fr giovanni.fusco@univ-cotedazur.fr karine.emsellem@univ-cotedazur.fr lucie.cadorel@inria.fr

Abstract. Real estate ads are a rich source of information when studying social representation of residential space. However, extracting knowledge from them poses some methodological challenges namely in terms its spatial content. The use of techniques from artificial intelligence to find and extract knowledge and relationships from textual data improves the classical approaches of Natural Language Processing (NLP). This paper will first conceptualize what kind of information on urban space can be targeted in real estate ads. It will then propose an automated protocol based on artificial intelligence to extract named entities and relationships among them. The extracted information will finally be modeled as RDF graphs and queried through GeoSPARQL. First results will be proposed from the case study of real estate ads on the French Riviera, with a focus on toponymy. Perspectives of quantitative spatial analysis of the geolocated RDF models of real-estate ads will also be highlighted.

Keywords: Geographic Information, Real Estate Ads, Social Representation of Space, Toponyms, French Riviera

1 INTRODUCTION

The retrieval and treatment of geographic information from textual data are a challenge in urban geography. New data sources are booming since the emergence of social networks, online platforms, etc. that offer the possibility to collect massive digital corpora. Indeed, *"a large number of statements that humans produce account for spatial phenomena"* [1]. This is the case of real estate ads. They are short texts presenting properties inscribed in urban space and are intended for a target population. They describe what are thought to be key characteristics of a piece of real estate, but they indirectly show how potential buyers/renters can project themselves in it, hinting also at the urban environment around it. They speak indirectly of individuals, social groups, places, urban forms, objects in space as they are represented by its authors who are trying to talk the symbolic language of their targeted readers. Real estate listings are not spatial reality, but a transformation of it for a particular purpose.

Thus, most of the real estate ad deals with the description of the property, anchored by observed and sometimes even measured elements (housing type, surface, number of rooms, quality of materials, energy efficiency, financial charges, price, etc.). When it comes to the location of the piece of real estate, its close or further urban environment, we wouldn't expect realtors to carry out any specific analysis: they will more or less consciously reflect the prevailing social representation of the neighborhood, eventually filtering negative aspects following conventional selling strategies. If we can extract the few elements concerning the spatial context of the property and if the ads can be geolocated, we have the possibility to project in space dominant aspects of shared social representations of different subspaces in the city. More generally, the analysis of a large corpus of geolocated real estate ads gives us the possibility to produce a geography of the social representation of urban spaces, and even to understand to what degree fragments of urban space become recognized and socially characterized places in the city [2]. Understanding real estate advertisements also makes it possible to grasp the strategies and behaviors of the various urban actors in space: inhabitants, real estate actors, developers, local authorities. The study of urban space through real estate ads requires the creation of an effective analysis protocol to be able to extract automatically and organize in a formal structure the socio-spatial phenomena included in the texts. There are many approaches and a variety of applications that allow processing and extracting information from textual data including geographic information. Various studies have already been initiated in this direction [3]–[6]. However, rare are those who ventured into the analysis of real estate ads and their peculiar characteristics.

The goal of this paper is thus to present a new methodological protocol, allowing to carry out studies on urban space from a corpus of geolocated real estate ads. The main challenge lies in the recognition, extraction, classification, and analysis of geographic information from the corpora of real estate ads.

The case study to which our protocol will be applied is a corpus of real estate ads from the housing market of the French Riviera (from 2019 and 2021). The proposed protocol will thus include a few specific steps linked to the syntax and the semantics of the French language. It is understood that every natural language will need a specific conceptualization of the way in which the valuable geographic information is encoded in the text of the ad. However, the overall structure of the protocol applies to any natural language and can open new perspectives in the use of real estate ads to understand the perceived geography of vast metropolitan areas.

The remainder of the paper will be organized as follows. In the next section we will highlight what kind of geographic information can be found and targeted in real estate ads. Section 3 will present a methodology to extract this geographic content, linking information extraction, relationship extraction and information modeling through RDF graphs. Section 4 will present first results from the analysis of real estate ads from the French Riviera and perspectives of future research.

2 GEOGRAPHIC INFORMATION IN REAL ESTATE ADS

Real estate ads are massive punctual data that contain descriptive texts partly focused on urban space. Real estate ads talk about space in different ways: they can describe physical features that are observable around the advertised property; they can describe the relational properties of the location by highlighting proximities and travel connections; they can give specific perceived attributes to the area, eventually using known toponyms; they can indirectly suggest what kind of populations are to be found in this spatial context. Of course, their readers are not geographers and planners. They will thus hardly use analytical knowledge or standardized representations of space (XY coordinates, census tracts, planning perimeters, location quotients, etc.). Even their use of isochrones (travel time) has to be understood as qualitative and evocative of lifestyles (what travel modes are implied? What destinations?) more than a precise description of urban space-time. They use more the evocative language of places than the abstract representation of space. These speeches of these places are to enhance the residential space and/or location in a particular urban context [3], [4]. Indeed, the literature indicates that location, or better socially perceived location, is a preponderant criterion in the acquisition of real estate: by choosing a piece of real estate, the buyer (or renter) chooses an address within a place, with all its implications in terms of aspired social status, lifestyle, and perceived needs within the life cycle of the household [5], [6], [7]. If social status, lifestyle and position within the life cycle of the household are the most important explanatory factors of residential mobility (together with ethnicity, as shown in [5]), we can assume that social representations of space carried by real estate ads will at least include these dimensions. We will thus expect to be able to identify places associated with higher vs lower social status, urban vs suburban lifestyles, needs for the young, the seniors, families with children, etc. But these associations will be rarely dealt with directly. Much more often they will have to be inferred by the semantic associations implied in the text.

Toponyms, i.e., place names, are specific issue of place analysis in geographic space through real estate ads. What we are interested in is the capacity of toponyms to structure the perception of places in the city. What subspaces are identified as places through the coherent use of a given toponym? What is the spatial extent of the use of this toponym? Are there areas where the use of toponyms is erratic or even absent? Can the use of a toponym beyond the area of its established perimeter be an indicator of the particularly strong social representation of the designated place? The use of toponyms within real estate ads will have to deal with specific disambiguation problems. If we don't want to pollute our use of real estate ads as a corpus of social representation of space in use, we have to consider toponyms, beyond their administrative definitions.

In order to take up these knowledge challenges, we must identify a few specificities of the description of space and place within real estate ads. These will call for specific methodological choices within our extraction and structuring protocol which we will present in the rest of the text.

First, as in any descriptive text for a generic public, the authors of real estate ads use common geographic entities to talk about places: city-center, neighborhood, village, sector, street, square, park, streetcar, train station, highway, school, campus, new project, etc are all common terms, well understood by people, beyond any formal definition by geographers or planners. It is also through the combination of these natural or man-made amenities, linear, punctual or surface-like features, that the place characteristics within which a property is located are evoked to the reader.

As we have seen, another fundamental category of geographic entities are the named entities, i.e., toponyms. These can come or not in association with common geographic entities. We can distinguish three situations:

- Geographic entity + associated toponym ("City of Nice", "Cimiez district", "Salis Beach"). These combined occurrences are particularly informative, because they directly associate the toponym to a more general category, indicating what the toponym means for the potential readers. And different geographic terms are differently evocative of social representations: identifying a given toponym as a *quartier*[neighborhood] evokes an urban life at a pedestrian scale or an older part of town, which is not the case if it is identified as a *secteur*[sector, district], which is a much more neutral subdivision of space.
- Unnamed geographic entity ("city", "neighborhood", "beach"). In this case the place is characterized as corresponding to a specific feature, but it appears nameless, hinting at a weakness of the placename and, possibly, at the whole place-making as a unique and specific location in urban space.
- Toponym alone without a corresponding geographic entity ("Nice", "Cimiez", "Salis"). This situation is the most enigmatic. We have evidence of the strength of a place name, but we don't exactly know what geographic feature is meant by that name.

Whenever placenames are used (cases 1 and 3), we can project their occurrences in space and study the spatial patterns resulting from them, eventually as a function of the associated geographic features.

Attributes associated to the geographic entities are also particularly informative of social representations. Indeed, the authors use the specific characteristics of places that give a context to the proposed property. Example: "famous district", "dynamic sector", "lively street", etc. Of course, the absence of any specific attribute could also be indicative of shared negative representations of a given place. We will thus never find descriptions like "crime-ridden district", "rough neighborhood", "dwindling village", etc.

Second, even if real estate ads mainly talk about places, they also use specific and relatively simple concepts of spatial relations: inclusion, proximity, concentration, connection, etc. Sometimes, even the multi-scale nature of property location within a neighborhood, a city and a whole metropolitan area is addressed. Spatial relations play a central role in the understanding of urban space (or spatial context of valorization). They make it possible to realize the relationships that the different geographic objects (the property, toponyms and/or common geographic entities) have in space. Examples: "In the heart of Cimiez", "Between the industrial zone and city-center", "Close to shops and transit", etc.

The words that express a spatial relationship are specific to the language of the real estate ads. As for French, two different ways of dealing with spatial relations arise:

- Simple spatial prepositions can encode the spatial relationship between two entities (E.g., in English, “in”, “on”, “at”).
- Compound prepositions that combine simple spatial prepositions with other words (nouns, adjectives, verbs) or prepositional locutions [1] that we call here “Space-time entities” (E.g., in English, “at 3 minutes from”, “near”, “close to”).

Beyond the specific linguistic encoding, there are four main types of spatial relationships in the texts of real estate ads:

- Spatial relationship of **situation**: We are talking here about location within a given space and at a given scale. Examples are: *“au coeur* [in the heart of]*”* represents location within the most central part of an entity, “*dans*[in]”, it represents insideness within a designated spatial extent, whether named or unnamed ; “*à*[at]” can also be used to represent location in a place, as if it were observed at a wider scale, where the place becomes a point on a map ; “*sur*[on]” and locates objects on topographic and hydrographic features. Sometimes, real estate ads omit the prepositions and the spatial relation with the entity and/or the toponym remains understood.
- Spatial relationship of **proximity**:
  - Temporal or metric distances that are quantities, real or perceived (E.g., “10 minutes from”, “100 meters from”).
  - Qualitative relationship of proximity (E.g., “close to”, “near”).
- Spatial relationship of **accessibility/connexion**: They are often complex since most accessibility relationships are between the places and not directly between the property and the places (E.g.: “Apartment near the station that **serves** Monaco”, “Neighborhood served by the line 4 of the subway”).
- Spatial relationship of **visibility**: this spatial relationship involves sight instead of movement. It completes our understanding of the spatiality of the geographic entities mentioned in real estate ads. Hence the interest in studying visibility relationships at the same level of more physical spatial relationships.

It is these different relationships that make it possible to understand the spatial context and the links between places. For example, the fact that the property is “in” the neighborhood of Cimiez or that the property is “close” to the neighborhood of Cimiez have a different meaning in terms of place-making. In the first case, we just remark the use of a given toponym to identify a place, which is already evidence of the toponym’s strength. In the second case, the strength of the toponym Cimiez is even higher: a piece of real estate which is out of its perceived boundary refers to the proximity to it in order to have a meaningful location for the reader. Conversely, the sense of place of the subspace where the property lies, whose toponym is not even mentioned in the ad, could be much lower.

In addition, the fact that the property is located “close” to certain features is informative of the targeted population and, therefore, of the kind of population associated with that space, in terms of all the relevant factors: social status, lifestyle, life cycle, ethnicity. Proximity to commerce and schools is not the same as proximity to a golf course, proximity to bars and nightlife is not the same as proximity to a park or to health-care facilities. What is to be considered here is the putative social status of the different facilities

(which could vary in different cultural contexts), the opposition between natural vs. urban amenities as well as the specific social groups (or just age groups) that a given facility or service caters to.

Third, when advertisers talk about features near the property, they also emphasize the presence of travel modes that allow readers to better understand the spatiality of the places and to project themselves into a wider urban space. However, the spatial relationships mentioned in the ads do not always allow assessing the precise distance between places. Two difficulties arise: The assessment of distance differs among people [8] especially in the context of real estate ads [9], and spatial relationships are often qualitative. What is more interesting is often the travel modes indicated (or just assumed) in the text, as it relates indirectly to lifestyles and/or to constraints in the household agenda. "Ten minutes from the motorway exit" is not just important for its quantification of the time-distance, but for its understanding of car mobility to access metropolitan space. "10-minute walk from schools and commerce", notwithstanding its truthfulness, hints at a pedestrian city-life or village-life. Indeed, according to the location factors cited, the spatial relationships expressed by the travel modes could also hint at different spatial scales and spatial contexts for a given place. For example, the fact that a property is at a particular walking distance from some geographic features contributes to the neighborhood characterization of the place, whereas its transit or car distance from higher-end services or centers could contribute to characterize a place as a cell within a larger urban and metropolitan space.

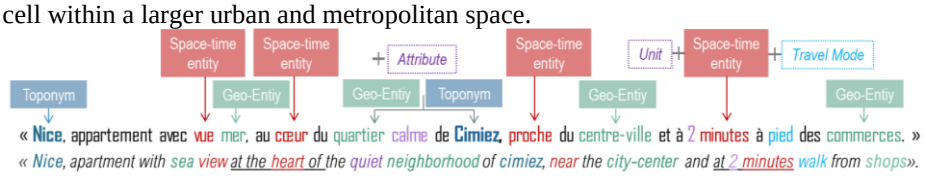

Fig. 1. Geographic information in the real estate ad

Figure 1 shows the main targeted features that must be extracted from a real estate ad to understand the perceived qualities of places in the city, with the subtleties of the semantic and spatial relations among them. In what follows we will see which protocols can be used to extract these features and to assign them the correct label while keeping the structure of the semantic and spatial relations.

3 METHODS TO EXTRACT GEOGRAPHIC INFORMATION AND RELATIONSHIPS

This section will deal with the specificities of automated processing of real estate ads to analyze social perceptions of urban space. To be able to use real estate ads, it is essential to transform raw ads into analyzable data which go beyond the usual pretreatments applied to texts such as data standardization, removal of special characters, lemmatization, and others. Usual techniques of lexicometry are of little use: pertinent information of social representation of urban space is minority within the real estate ad and can be scattered within it. Thus, analysis of associations based on word co-

occurrences could even be misleading [10]. We need to know what terms are used to describe the piece of property and which ones are used to characterize its spatial setting. An understanding of the language syntax is thus necessary prior to lexicometry.
The challenge is therefore based on the creation of an adapted protocol to deal with the socio-spatial phenomena and geospatial information in the texts. There are different methods and techniques that can be used to process texts, including geographic textual information [11], [12]. To detect and extract the targeted geographic knowledge in the real estate ads, we decided to implement a particular protocol based on NLP (Natural Language Processing) techniques. This protocol can only be realized with a hindsight and conceptualization of the knowledge present in the texts and the phenomena to be described. It is therefore necessary to know beforehand a set of indicators that provide information as presented in the previous section.
The extraction of information from texts can be established on specific methods using artificial intelligence techniques. The chosen method is based on the extraction of all the information and its relationships in two steps: The named entities recognition (NER) and dependency parsing.

3.1 INFORMATION EXTRACTION (IE)

The Information Extraction (IE) from text is made possible, among others, by the named entities recognition (NER) approach. NER is based on a natural language processing (NLP) technique that makes it possible to extract valuable elements from a text in particular by "*the automatic identification of the entities present in a text to the classification of these entities into different categories*" [13]. This method is among the most widely used methods for extracting a mass of information automatically from text [14], [15]. It is based on machine learning approaches. The real challenge here is to be able to extract all the geographic information automatically in the rich corpora of real estate ads.

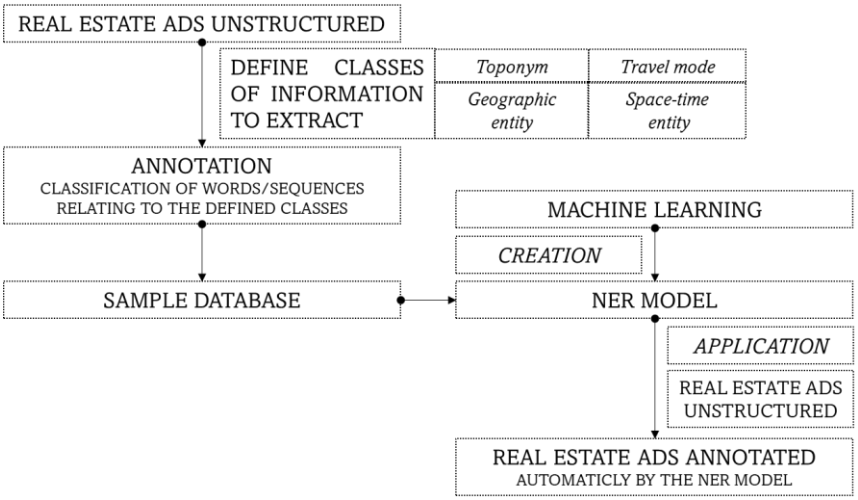

Fig. 2. Steps to extract information from the real estate ads by using NER Model.

Figure 2 shows the different steps of the method that will be detailed below:

The **first step** is focused on the identification of the targeted information, i.e., the geographic information previously presented that echoes the environment of the property: geographic entities of several kinds, toponyms, travel modes, and the words used to create a spatial relationship (space-time entities).

The **second step** creates a sample database of real estate ads annotated according to the previously defined classes (fig.3). The goal is to have a training dataset for the creation of the NER learning model. For this, it was necessary to classify manually the words of each category identified in each real estate ad. The annotation is a very time-consuming task that must be performed on a significant number of real estate ads for the training dataset to be effective. In this research, we annotated more than 1000 ads of the real estate market of the French Riviera. They were written by professional real estate agents in 2021 and concern all types of real estate (sale and rental; apartments, houses and garages). This sample is representative of the real estate market of the French Riviera in its diversity (coastal cities, suburban areas, hinterland, etc.).

Fig. 3. Annotation of real estate ads according to the information selected.

The **third step** is based on the training of a NER learning model. There are many NER learning models [16], [17]. SPACY and FLAIR are two leading models for performing the named entity recognition task. These models are different in terms of neural network architecture and algorithms, requiring different computing power. In this study, we have tested these different NER models on our annotated French real estate ads. The results and performances have been detailed in a previous paper [18]. Here we will focus on FLAIR which is based on the BiLSTM Language Model. The advantages of this model are the consideration of the context of labeled words and the fact that it is more easily optimized [19].

At the end of the information extraction task, we obtained annotated ads according to the defined classes and a list of words (all named entities) for each ad. The extracted elements of information have no links among them, and the geographic information is devoid of context. In order to reconstruct context and relations, we need to go through automatic syntactic analysis, as shown in the next section.

3.2 RELATIONSHIP EXTRACTION (RE)

Methods of Relationship Extraction (RE) pay attention to the semantic structure of the text and are based on syntactic analysis of sentences or groups of words that compose the text [20]. We selected the dependency parsing method deriving the links between words from the lexical relationships that words keep in the text. The objective here is to extract, from the grammatical dependencies, all the links between geographic objects and their socio-spatial properties through the lexical and sometimes specialized relationships that animate them. There are different libraries to study the grammatical

relationships of texts. We chose the STANZA library developed by Stanford NLP Group [21] and which adapts to French texts. This library allows to highlight the grammatical structures between the words in the text by *"creating a tree of words from an input sentence that represents the syntactic dependency relationships between words"* [21] (fig.4). The dependency tree was created from the structure of sentences and the nature of words (called Part-of-Speech (POS)). Stanza's dependency tree of texts makes it possible to understand the grammatical relationships between each word or sequence of words for each sentence of each real estate ads (fig.4).

The careful analysis of grammatical dependencies of real estate ads, more particularly by focusing on the lexical links among named entities make it possible to highlight four main types of relationship: association, spatial, modal, and attribute relationship, as detailed below.

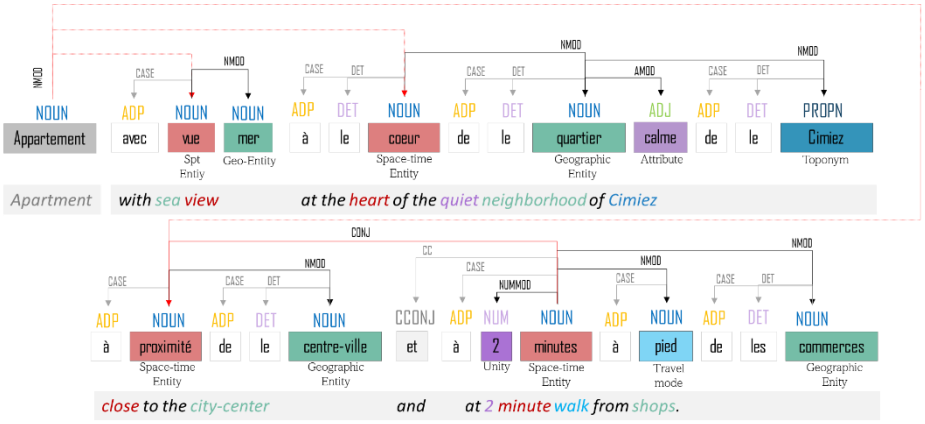

Fig. 4. Grammatical dependencies of STANZA on the real estate ad.

Table 1. List of Part-Of-Speech (POS) of the tokens in the ad (fig.4).

Part-of-Speech (POS)	*Noun (NOUN), Adposition (ADP), Determiner (DET), Proper noun (PROPN), Adjective (ADJ), Coordinating conjunction (CCONJ).*

Table 2. STANZA legend of the dependency tree in the ad (fig.4).

TYPE	**LEGEND OF DEPENDENCIES [21]**
[NMOD] Nominal modifier	*« Is used for nominal modifiers of nouns or noun phrases. »*
[AMOD] Adjective modifier	*« Is any adjectival phrase that serves to modify the meaning of the nominal head. »*
[NUMMOD] Numeric modifier	*« Is any number phrase that serves to modify the meaning of the noun with a quantity. »*
[CONJ] Conjunction	*« Relation between coordinated elements. The head of the relation is the first conjunct and other conjuncts depend on it. »*
[CC] Coordinating conjunction	*« Relation between the head conjunct of a coordinate structure and any of the coordinating conjunction involved in the structure. »*
[CASE] Case	*« Is used for any preposition introducing a nominal construction in French. Prepositions are treated as dependents of the noun … »*
[DET] Determiner	*« Relation between the head of a nominal phrase and its determiner. »*

Association relationships focus on the dependency between two kinds of named entities (toponyms and geographic entities) to recompose the geographic objects. This makes it possible to understand exactly which toponym is related to which geographic entity in the texts. The relationship between toponyms and a geographic entity is usually a direct nominal relationship (NMOD) as can be seen in this example (tab.2). E.g., Relation between the geographic entity "*quartier*[Neighborhood]" and the toponym "Cimiez" (fig.4).

Spatial Relationships express the spatial context of the geographic objects, the spatial scale of valorization, the surroundings of the property, the position that a geographic object takes in space. There are different spatial relationships in these texts between the property and the geographic objects. The grammatical relationships as spatial relationships are rarely direct between the property and geographic objects. Spatial relationships are expressed by prepositional locutions based on annotated space-time entities (e.g., "Apartment close to the city-center", "Apartment at two minutes from shops"). The links are only made through a space-time entity (*proche*[near], *minutes*[minutes], *proximité*[close] (fig.4)). The space-time entities that are at the center of these spatial relationships are between the property and geographic objects. In this case, the property has a direct nominal modifier relationship (NMOD) with a space-time entity and the latter has a second grammatical relationship with a geographic object (fig.4). Exceptions exist when geographic objects are simply cited or when the spatial relationship is based on spatial prepositions considered as an adposition (ADP) (e.g., "Apartment *dans*[in] Nice", "Apartment *sur*[on] the mountain", "Nice"). The spatial relationship based on the spatial preposition is directly recognized in the grammatical dependencies because the property has a direct grammatical link with a geographic object as a nominal relationship (NMOD).

Modal Relationships concern the travel modes when they express a distance between the geographic objects (either between the property and the geographic object or between two geographic objects). Generally, the travel mode has a nominal modifier relationship (NMOD) with a named entity. In this example (fig.4), the travel mode "*pied*[walk]" has a grammatical relationship with a space-time entity "*minutes*[minutes]" making the spatial relationship between the property (subject) and the geographic entity "*commerces*[shops]".

Attribute relationships provide a quality of the geographic objects mentioned in the texts (E.g., "Dynamic sector", "Quiet district"). These characteristics are an additional key to understanding spaces in their social dimension. Generally, the named entities have a direct adjective modifier relationship (AMOD) with their characteristics because these are most often adjectives (E.g., relationship between "*quartier*[neighborhood]" and "*calme*[quiet]" (fig.4)). In addition, spatial relationships can also have attributes (E.g., qualitative attribute "very close to" or quantitative attribute "2 minutes from" where "2" have a numeric modifier relationship (NUMMOD) with "minutes" (fig.4)).

The extraction of these relationships is founded on a rule-based approach according to recurring identified structures expressed by the grammatical relationships. Many parameters have been added to integrate the specificities of each relationship (number of words between tokens, type of part-of-speech of token, type of relationship between tokens, etc.). This step makes it possible to link all the pieces of geographic information together and give them meaning and context.

To conclude, the extracted information and relationships can be modeled in the form of a grammatical dependency graph. A dependency graph is a set of dependency relationships between tokens in a sentence [22], where the tokens can be represented by a node and the dependencies by arcs between nodes. In our case, the nodes are the pieces of extracted geographic information (space-time entity, toponym, geographic entity, travel mode and attribute) and the arcs are the grammatical relationships extracted among them (spatial relationship, association relationship, attribute relationship, modal relationship) (fig.5).

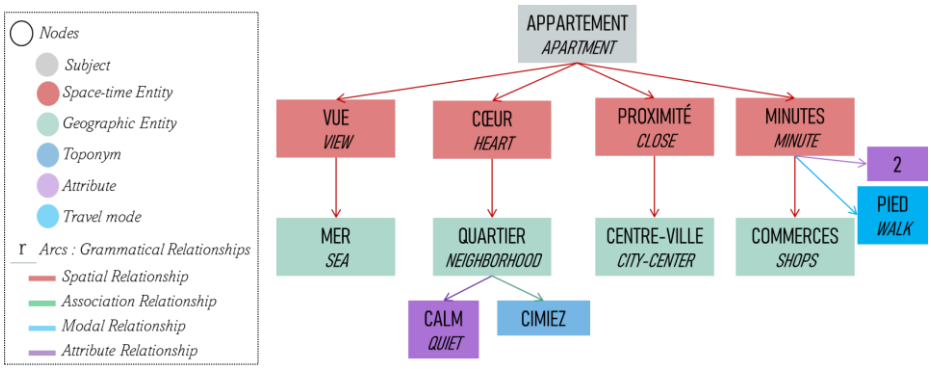

Fig. 5. Conceptual grammatical dependency graph.

3.3 STRUCTURING INFORMATION EXTRACTED

The analysis of the extracted geographic information is only possible after the transformation of the real estate ads into analyzable data. The objective is to structure the extracted entities and their relationships in a common graph structure. Several modeling choices allow us to produce a common graph structure for all the ads.

First, we include white/empty nodes in the model to indicate the absence of some key pieces of information (fig.6). In fact, the real estate ads are not constituted in the same way: some do not refer to the geographic area where the property is situated, others do not mention any toponym, some mention just the geographic entities without associated toponyms. The presence of empty nodes allows all real estate ads to keep the same structure but also to model the absence of information.

Second, we link all information related to the geographic object to toponyms, whether they are present or not in the ad (fig.6). A toponym can be considered as a unique identifier even if it is linked to several types of geographic entities. We will therefore link the geographic entities to the toponym. The possibility of linking an

attribute node to the toponym (and, through the latter, to the geographic entity) is always foreseen. These nodes can, of course, be empty (fig.6).

Third, we link the property to the toponym in order to represent the spatial relationship between them. All the information relating to the kind of spatial relationship that has been extracted are modeled as parameters of the link (presence or not of space-time entity, travel mode, and attribute) (fig.6).

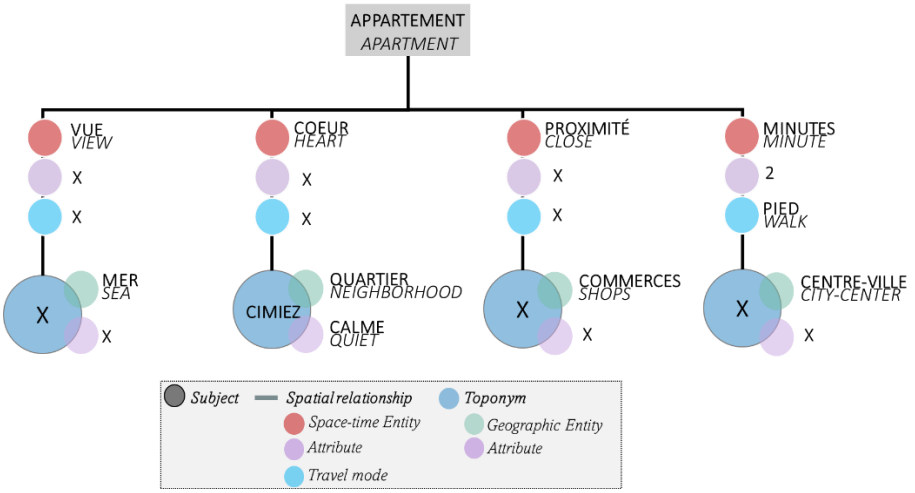

Fig. 6. Ad structured according to the extracted geographic information and relationships

Figure 6 gives an example of the resulting graph. In this ad, a piece of property (here an apartment) has a spatial relationship with four geographic objects:

- The first spatial relationship is with an "empty" toponym concerning the "*mer*[sea]" as geographic entity through a visibility relationship according to the space-time entity "*vue*[view]". In many ads of the French Riviera, the toponym "*Méditerranée*[Mediterranean]" is understood for the geographic entity "*mer*[sea]".
- The second spatial relationship is of situation kind, according to the space-time entity "*coeur*[heart]" with the toponym "Cimiez" which is considered in this ad as a "*quartier*[neighborhood]" that is "*calme*[quiet]" according to its attribute.
- The third spatial relationship is one of distance between the apartment and the "*centre-ville*[city-center]" through the space-time entity "*proximité*[close]". We do not have more precision for this spatial relationship because of the absence of information at the level of the parameters "travel mode" and "unity" or "attribute";
- The last spatial relationship is also of distance between the apartment and the "*commerces*[shops]" according to the space-time entity "*minutes*[minute]". We can see that its attribute is "2" and its travel mode is "*pied*[walk]".

Structuring ads, first conceptually and then practically in a graph, is of paramount importance to make their exploitation possible. The transformation of real estate ads also allows to store them properly, for example, in the form of an RDF graph. The RDF graph is an appropriate formalization for transformed real estate ads since relationships

are grammatical dependencies and follow the same logic as the grammatical dependency graphs that we created. The basic components of RDF data modeling are a triple of elements: a first node (the subject), a second node (the object) and a relation between the two (the predicate). This intrinsic graph-based framework can also be easily queried. Its GeoSPARQL vocabulary is particularly well adapted to geographic information and makes it possible to make spatial queries, opening the field of possible applications for the analysis of urban space from real estate ads.

4 FIRST RESULTS FROM THE FRENCH RIVIERA AND PERSPECTIVES

Whenever they contain precise location (XY coordinates or street address), real estate ads become punctual data and therefore give the opportunity, once transformed, to carry out different spatial analyses to understand urban spaces in their socially perceived dimension. In this paper, we will limit ourselves to different cartographic representations of specific query results from the presented protocol. Their spatial analysis through appropriate algorithms will be a further research phase.

Application 1: The spatial distribution of the real estate ads that use the toponym "Gambetta" in the city of Nice (France) within a corpus of real estate ads geolocated at the address within the city of Nice in 2019.

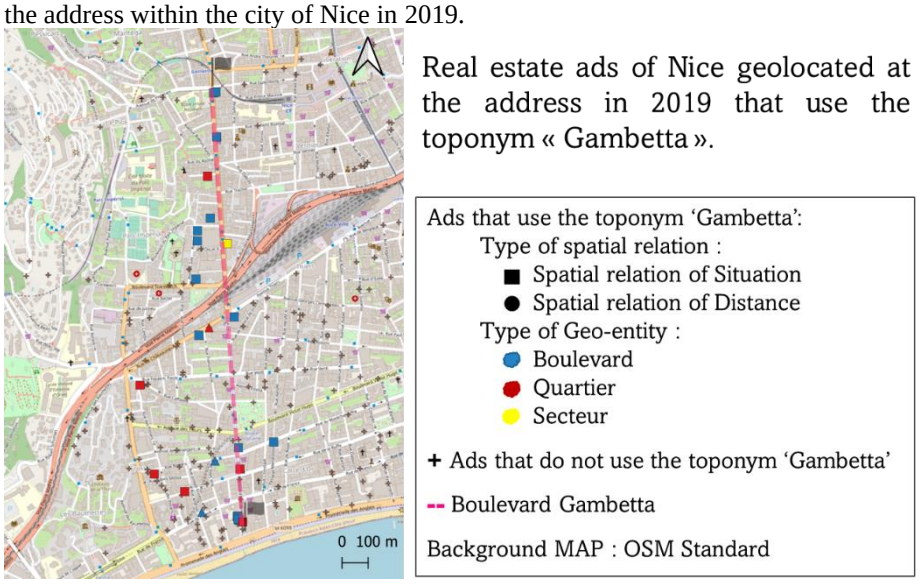

Fig. 7. Real estate ads of Nice geolocated in 2019 that use the toponym "Gambetta".

Many real estate ads mention the toponym "Gambetta" which makes "Gambetta" a frequently used toponym in the marketing of real estate ads in our study area. There are different geographic entities related to the toponym "Gambetta". In the real estate ads in Nice, the toponym "Gambetta" can be of the "Boulevard" type which makes it an odonym but can also be used to designate a "neighborhood" and a "sector" (fig.7). The

widespread use of the placename Gambetta makes it a prominent place within the social representation of the city of Nice, insisting on its urban connotation: a main street (boulevard) and a neighborhood, the latter being used mainly for its southern section. The more generic and less urban word "sector" is seldom used, and our relatively limited sample doesn't contain other designations of the street denoting its role of transportation artery. Further analyses of this point pattern are needed on a larger corpus of data, and in conjunction with the attributes associated to the boulevard/neighborhood/sector, as well as with the other geographic features included in the RDF graphs, the associated transport modes, etc.

Application 2: The spatial distribution of the real estate ads that use the toponym "Promenade des Anglais" in the city of Nice (France) within a corpus of real estate ads geolocated at the address within the city of Nice in 2019.

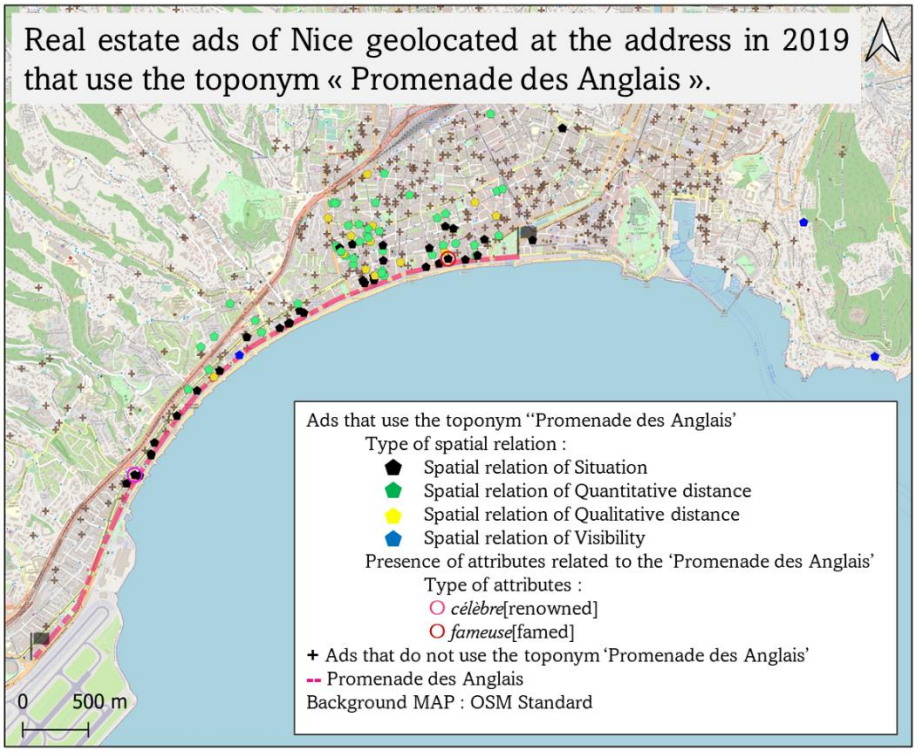

Fig. 8. Real estate ads of Nice geolocated in 2019 that use the "Promenade des Anglais".

The Promenade des Anglais is one of the most mentioned places in real estate ads in Nice in 2019. Some ads locate the property by the proximity to the Promenade des Anglais using a quantitative or a qualitative distance while others use the notion of visibility (fig.8). This toponym has thus an area of influence beyond the perimeter of properties situated on it (at least according to the ads). Even the "view" of the Promenade des Anglais is a very rewarding criterion. Its view from the eastern hills is a "canonical view" within the imagery of the French Riviera as painted or photographed by

artists and then used by tourist guides and other marketing supports for almost two centuries [23]. Interestingly enough, the Promenade des Anglais has an area of influence well beyond the coastal strip, but is not used to characterize the situation of real estate ads in its westernmost section. We observe here a clear discrepancy between the official odonym and the social representation of the place Promenade des Anglais. For the users of real estate ads, the Promenade des Anglais is the famous seafront promenade of the city of Nice. Some ads develop this dimension of social representation by the presence of attributes such as "*célèbre*[renowned]" or "*fameuse*[famous]". When the street departs from the seaside and is bordered by the wastewater treatment plant and the airport, its surrounding areas are no longer considered part of this well-established urban place.

Application 3 : The spatial distribution of the real estate ads that use the toponym "Montfleury" in the city of Cannes (France) within a corpus of real estate ads geolocated at the address within the city of Cannes in 2019.

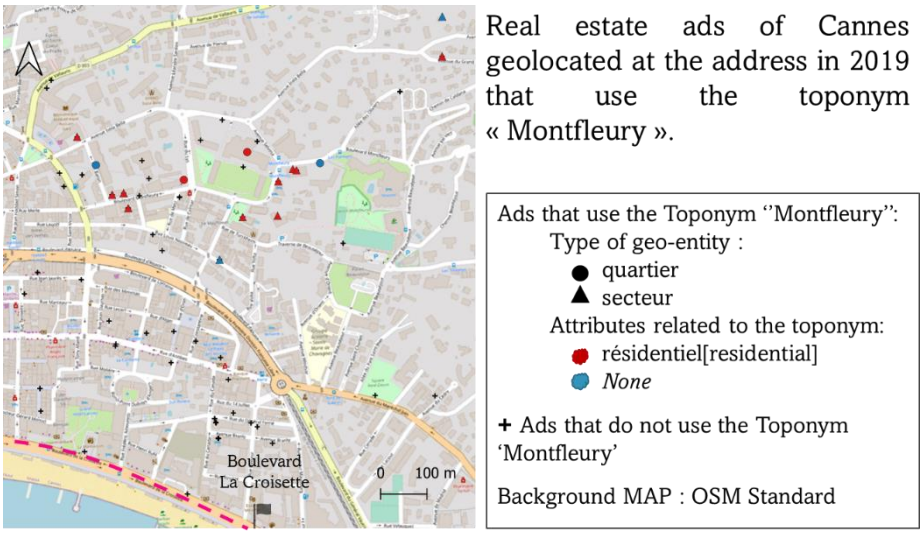

Fig. 9. Real estate ads of Cannes geolocated in 2019 that use the toponym "Montfleury".

According to the number of real estate ads located at the address in 2019 within the city of Cannes, the real estate market in "Montfleury" was very dynamic (fig.9). We would therefore like to know from the real estate ads what are the characteristics attached to the toponym "Montfleury". In fact, for more than half of the real estate ads mentioning the toponym "MontFleury"/"Montfleury", this is considered as "Residential" and only residential (no other attributes are used within the ads). Residential areas are less generally considered neighborhoods in France and Montfleury is not an exception. The term *"secteur"*[sector] is thus much more frequent than *"quartier"*[neighborhood], in contrast with what was observed for the toponym Gambetta in Nice.

Beyond these three cartographic analyses, we can easily set the goal of conducting quantitative spatial analysis on this data to provide an additional key to understanding

the social representation of urban spaces through real estate ads. Different statistical methods and spatial analysis techniques of point patterns analysis (PPA) could be applied to the geolocated ads and their associated RDF graphs. For example, it would be very interesting to work around the question of toponymy (dominant and most used toponyms within a given study area) while integrating space (intensity of use of toponyms in space, spaces that have a hesitant toponymy when there are different toponyms to designate the same space, spaces that have a low of presence of toponyms, etc.). These spatial distributions of the punctual data can be measured by applying methods based on spatial density (kernel density, DBSCAN) and by using methods around the fuzzy logics, to model core and support space for a given toponym. Area of influence of toponyms could also be studied beyond the relationship of spatial situation, as shown in the example of the Promenade des Anglais in Nice.
We could also study the factors of location used within the real estate ads located in a given space (type of geographic object mentioned, multiplicity of the factors of location or not and their copresence in an ad and in all ads, preponderant characteristics mentioned, as well as absence of the characteristics). We can show the consensus of information if it exists through the study of convergences and divergences within the same space, and infer the target population of the ads (if they are not explicitly mentioned). Linking target populations and mentioned places will be another important aspect of their social representation.
We can also study the question of the reputation of places by studying the characteristics associated to the places mentioned in the real estate ads (dominant attributes, multiplicity of attributes to qualify a given place, copresence of attributes, absence of attributes, sense of attributes).

These perspectives will be the object of a further research endeavor. We think that the main contribution of the present paper is to have proposed a new methodology for extracting spatial information from real estate ads, that opens the way of spatial quantitative analysis of one of the most qualitative aspects of urban research: social representation of places.

Acknowledgement*: This research was carried out thanks to a research grant by KCityLabs, KINAXIA Group (CIFRE Agreement with UMR ESPACE).*

References

1. Stosic, D. : « ‘par’ et ‘à travers’ dans l’expression des relations spatiales : comparaison entre le français et le serbo-croate » (2002). Available: https://hal.archives-ouvertes.fr/tel-00272907/
2. Relph, E.: « Place and placelessness » (1976). doi: 10.4135/9781446213742.n5.
3. Alba, M. et al. : « La publicité immobilière à l’assaut de l’environnement dans une grande ville du Sud, Mexico, 1950-2000 ». Ecol. Polit., vol. N°39, no. 1, p. 55. (2010). doi: 10.3917/ecopo.039.0055.
4. Blanchi, A. et al.: « The real estate ads, a new data source to understand the social representation of urban space » ECTQG21, (2021).

5. Shearmur, R. et al.: « From Chicago to L.A. and back again: A chicago-inspired quantitative analysis of income distribution in montreal ». Prof. Geogr., vol. 56, no. 1, pp. 109–126 (2004). doi: 10.1111/j.0033-0124.2004.05601016.x.
6. Thomas, M.-P. : « Les choix résidentiels : Une approche par les modes de vie ». pp. 1–41 (2018).
7. Sigaud, T. : « Accompagner les mobilités résidentielles des salariés : l'épreuve de l'entrée en territoire ». Espaces et sociétés, 162, 129-145. (2015).
8. Bailly, A. : « Ditances et espaces : vingt ans de géographie des représentations ». Espac. géographique, vol. 14, no. 3, pp. 197–205 (1985). doi: 10.3406/spgeo.1985.4033.
9. McKenzie, G. et al.: « The 'Nearby' Exaggeration in Real Estate ». Proc. Cogn. Scales Spat. Inf. Work. (CoSSI 2017), (2017).
10. Lancia, F.: « WORD CO-OCCURRENCE AND SIMILARITY IN MEANING: Some Methodological Issues » Mind as Infin. Dimens., pp. 1–39, (2007).
11. McKenzie, G. and al. : "Identifying urban neighborhood names through user contributed online property listings," ISPRS International Journal of Geo Information, vol. 7, no. 10. (2018).
12. Hu, Y. and al. : "A Semantic and Sentiment Analysis on Online Neighborhood Reviews for Understanding the Perceptions of People toward Their Living Environments," Ann. Am. Assoc. Geogr., (2019)
13. Shrivarsheni,: « How to Train spaCy to Autodetect New Entities (NER) » (2020). Available: https://www.machinelearningplus.com/nlp/training-custom-ner-model-in-spacy/
14. Andrey from Prodigy Support: « Former ensemble NER et extraction de relations (RE) ». pp. 3–5, (2021). www.support.prodi.gy/t/training-ner-and-relations-extraction-re-together/3911
15. Wang, J. et al.: « NeuroTPR: A neuro-net toponym recognition model for extracting locations from social media messages ». Trans. GIS, vol. 24, no. 3, pp. 719–735, (2020). doi: 10.1111/tgis.12627.
16. Benesty, M.: « NER algo benchmark: spaCy, Flair, m-BERT and camemBERT on anonymizing French commercial legal cases ». Towar. Data Sci. (2019). Available: https://towardsdatascience.com/benchmark-ner-algorithm-d4ab01b2d4c3
17. Hu, Y. and al.: « How Do People Describe Locations during a Natural Disaster: An Analysis of Tweets from Hurricane Harvey » Leibniz Int. Proc. Informatics, LIPIcs, vol. 177, no. 23, pp. 1–16, (2020).
18. Cadorel, L. et al.: « Geospatial Knowledge in Housing Advertisements: Capturing and Extracting Spatial Information from Text » (2021). HAL Id : hal-03518717.
19. Duffy, S.: "Is Flair a suitable alternative to SpaCy?" (2020). https://medium.com/@sapphireduffy/is-flair-a-suitable-alternative-to-spacy-6f55192bfb01
20. Perera, N. et al.: « Named Entity Recognition and Relation Detection for Biomedical Information Extraction ». Front. Cell Dev. Biol. 8.(2020).
21. Sanford NLP Group, « Stanza - A Python NLP Library for Many Human Languages | Stanza ». Available: https://stanfordnlp.github.io/stanza/ | https://universaldependencies.org/

22. Alfared, R. :« Acquisition de grammaire catégorielle de dépendances de grande envergure » (2013). HAL Id : tel-00822996.
23. Hérault, M. « La Riviera, pays de l'éternel printemps : Imaginaire paysager et transferts culturels, à Nice et dans son territoire, du Grand Tour à nos jours », Thèse de Doctorat, Sorbonne Université, Paris (2021) https://www.theses.fr/2021SORUL022