University of Zurich^UZH

Identifying toponyms and location references in residential real estate listings in Zurich City

GEO 511 Master's Thesis

Author
Chantal Angela Meier
15-739-915

Supervised by
Prof. Dr. Ross Purves

Faculty representative
Prof. Dr. Ross Purves

01.10.2023
Department of Geography, University of Zurich

Acknowledgments

I would like to express my gratitude to the following individuals, without whom my journey would have been considerably more challenging:

First and foremost, I would like to express my sincere thanks to Prof. Dr. Ross Purves. His unwavering guidance, invaluable support, and countless useful tips throughout the conduction of this thesis have been instrumental in shaping my understanding and success. Above all, his encouragement has been a constant source of motivation for me.

I must also extend my heartfelt thanks to Dr. Daniel Sager, owner of Meta-Sys AG, for his pivotal role in making this project a reality. His generosity in providing his dataset on residential real estate listings was absolutely indispensable, as without it, the comprehensive analysis would not have been possible. I am profoundly grateful for his willingness to share his valuable resources.

I would like to extend a special thank you to my dear long-term smartest companion and most impressive beard wearer Cedric Kaissl, for coming to my aid with the code. His advice on my questions has been a significant factor in the success of this project.

Iva Pavlicevic, another dear friend, deserves a heartfelt thank you for the countless productive working sessions that helped me stay focused and on track and for proofreading the chapter on data and methodology.

I am also indebted to my father Felix Meier for his generosity in lending me his laptop during the whole time, when mine struggled to process the vast amount of data required for this project.

I would like to acknowledge my roommates and friends, Bastian Gruber, Livia Thoma and Noa Rieger, who were an important pillar. Noa, your encouragement, support, and surprises with caffeinated drinks and dextrose have been invaluable. Livia, thank you for your kind words and for taking on some household tasks during my moments of urgency.

I extend my sincerest thanks to each of you for your contributions, kindness, and support. Your assistance has played a significant role in my success, and I am profoundly grateful.

Abstract

Naive geography, and vernacular geography as a subset of it, are crucial concepts that delve into human perceptions of the spatial environment. This knowledge is accumulated over a lifetime and is inherently extensive for places where individuals reside or spend prolonged durations. Vernacular geography primarily concerns itself with places and spatial relationships. Such places are often termed as "fuzzy" places or toponyms since their boundaries, unlike administrative units, are indistinct. For instance, where precisely does the "Midwest" lie? Similarly, spatial relationships are not explicitly quantifiable: what exactly does "nearby" imply?

In human-to-human communication, such vague concepts generally pose no challenges since we intuitively grasp and interpret them. However, this is not the case in human-machine interactions. An example can be seen in web search queries, which have popularized information extraction. Most search queries encompass a spatial component, vital to our daily activities. Thus, studies aimed at better understanding vernacular toponyms and spatial expressions are essential to enhance the efficiency of human-machine interactions.

Understanding vernacular toponyms and spatial relation expressions is a core focus of Geographical Information Retrieval (GIR), an extension of classic Information Retrieval. Central processes in this field include Toponym Recognition, which detects place references from unstructured sources, typically text, and Toponym Resolution, where identified toponyms are mapped to specific places.

For this thesis, named entity recognition is conducted using the freely available spaCy model to detect place references in a dataset of residential property listings in Zurich. The identified locations are subsequently mapped and spatially analyzed using kernel density estimation. The analysis revealed that the most commonly used place references pertain to generic location descriptions (such as 'central' or 'quiet' locations), significant landmarks (transport hubs or places of high renown), natural landmarks like bodies of water and mountains, as well as well-known neighborhoods and public squares.

The spatial analysis indicated that certain prominent terms are used excessively, resulting in a lack of discernible spatial pattern, as they appear ubiquitously across the entire urban area. In contrast, other terms allowed for the analysis of the perimeter within which a place or transport hub is deemed significant, the perceived proximity to specific sites, or viewpoints from where certain landmarks, like the Alps, can be observed.

Contents

List of Tables

List of Figures

Notation

Abbreviation	**Meaning**
AI	Artificial Intelligence
FSO	Federal Statistical Office (of Switzerland)
GIR	Geographic Information Retrieval
GIS	Geographic Information System(s)
IE	Information Extraction
IR	Information Retrieval
KDE	Kernal Density Estimation
MUC	Message Understanding Conferences
MUC-6	Sixth Message Understanding Conference
MUC-7	Seventh Message Understanding Conference
NER	Named Entity Recognition
NLP	Natural Language Processing
OSM	OpenStreetMap
POI	Point of Interest
SRE	Spatial Relation Expressions
UGC	User generated content
VGI	Volunteered Geographic Information

1 Introduction

1.1 Context

The knowledge of the local population about their surroundings has been captured since the 1990s with the term "naive geography" introduced by Egenhofer & Mark (1995). A significant subset of this is vernacular geography, which deals with spatial relationships and toponyms of vague geographical extent, so-called fuzzy places (Montello et al., 2003).

These vernacular geographies are of daily significance because people speak and think within them, and typically understand exactly what is meant, even in the absence of clear definitions (Davies et al., 2009; Montello, 2009). Fuzzy places and our knowledge about them also exert psychological influences, as they dictate, for instance, which areas we visit for certain activities or the times at which we avoid others (Evans & Waters, 2007).

However, these vague concepts are not as trivial to interpret for machines as they are for us humans (Lieberman & Hanan, 2011). Thus, they are often overlooked in official gazetteers and GIS systems because, on the one hand, they are not widely recognized, and on the other hand, they require specialized solutions since our systems are designed for sharp delineations and precise coordinates. (Jones & Purves, 2008).

Nevertheless, these vernacular geographies are of paramount importance when it comes to improving search queries or, more broadly, the interaction between humans and machines. This is because the vast majority of queries on search engines or similar platforms include a spatial component, and due to the lack of such vernacular toponyms and spatial relationships in the used gazetteer, they can only be inadequately addressed or require multiple queries by the user to be served correctly. (e.g. Jones & Purves, 2008)

With the growth of the internet, the need for a better implementation of this knowledge in various search engines has not only increased (e.g., How should a search engine specifically interpret when someone searches for a coffee "nearby" or "downtown"?), but so have the sources from which information can answer questions about spatial relationships and vernacular toponyms.

Volunteered Geographic Information (VGI) plays a pivotal role in this context, having proliferated concurrently with the advent of Web 2.0. (Pasley et al., 2008). This includes voluntarily generated user content with a spatial reference, as found in travel blogs, georeferenced images or tweets, real estate listing descriptions, and many more.

This is where the field of GIR (Geographic Information Retrieval) comes into play, an extension of the classical IE (Information Extraction). Its primary objective is to extract geographical information from various sources with methods as natural language processing and named entity recognition (NER) and associate these toponyms with a location in the real world. This specifically involves the processes of toponym recognition and toponym resolution, collectively referred to as geotagging.

1.2 Research gap

To date, there are limited studies that utilize real estate data beyond a property-related context, especially not as a source for spatial references. Yet, such data inherently offers immense potential for GIR. These datasets are widespread and comprehensive, while many social media platforms tend to focus primarily on "tourist hotspots". Furthermore, the location of a property is one of its most vital attributes, resulting in a strong incentive to provide extensive descriptions. (Hu et al., 2019)

In addition to the utilized datasets, the geographical location (Zurich/Switzerland) and the language of the source (German) can also be identified as research gaps or at least as areas where not much research exists. In general, there are few studies employing natural language processing for the German language, with the majority of research existing for English. There are scant GIR studies focused on Switzerland or Zurich. Notable exceptions include Hollenstein's analysis of Flickr images in Zurich's city center (Hollenstein, 2008) and the Text+Berg Corpus, which has been studied for Switzerland (Derungs & Purves, 2014; Ettlin, 2011).

The last but not least research gap addresses the approach conducted in this thesis. Many studies don't systematically search for all possible location references within a given geographic area. Instead, they often adopt the reverse approach, where they search for a specific set for toponyms or spatial references. An exception can be found in the study of Twaroch et al (2008), where they searched vernacular toponyms within the social website of Gumtree.

1.3 Aims and research questions

This master's thesis aims to extract as many location references as possible out of real estate listings through a natural language processing, or more specifically a named entity recognition (NER) model called spaCy. The used dataset consists of crawled real estate listings for Zurich City, covering a long time period from 2004 to 2022.

In a subsequent step, these perceived vernacular geographies can be cartographically represented through the available geocoordinates assigned to the address of the listings and then spatially analyzed through Kernel Density Estimation.

Given the size of the dataset (both in terms of the number of listings and geographical coverage), these shall hopefully allow conclusions regarding which location references are deemed significant by the local population.

If applicable, they can be compared to official toponyms and their locations or, in the case of polygons, their boundaries. Any discrepancies observed can then be discussed.

Specifically, the following research questions are addressed:

1. *Which location references are of utmost relevance in Zurich-City?*
2. *What spatial distribution patterns are typical for these location references, how do they possibly differentiate from official boundaries and locations, and how can they be explained?*

1.4 Structure

To address the aforementioned research questions, this thesis adheres to the conventional structure of Background, Data, Methods, Discussion, and Conclusion.

The Background illuminates the relevant literature associated with the addressed concepts, specifically naive and vernacular geography, and their significance to geography and other fields. The ensuing section delves into the practical application of these concepts in the fields of GIR. This part elucidates the core tasks and challenges of GIR (toponym recognition, toponym resolution, and toponym disambiguation) in detail and contextualizes GIR as a whole field, distinguishing it from the older domain of Information Extraction (IE). Special attention is accorded to the processes of toponym recognition using methods of Named Entity Recognition (NER). Subsequently, a brief overview of gazetteers and corpora, as important sources of GIR are explained. There's an in-depth explanation of volunteered geographic information (VGI), and within that, the data from real estate listings. The final part of the Background offers an overview of the Zurich real estate market, as understanding it is essential to better interpret the data generated from it, upon which the present analysis is based.

The Data section provides a snapshot of the dataset, its composition, coverage, and how it has been processed and pre-processed by the manufacturer, Meta-Sys. A brief overview of other used data, like official statistics on Zurich's real estate market from federal and city sources, is also provided.

The Methodology explains how the data is validated for its extent by comparison with official statistics and is cleaned of duplicates. It then elucidates the core task - the NER conducted with spaCY, and the subsequent enhancement of this model, including the addition of n-grams to refine the "location" concept until sufficient precision and recall for the extracted location references are achieved. In the final step, the chosen method of Kernel Density Estimation (KDE) for spatial analysis is explained.

The Results section offers insights into the extracted location references, which are classified into various categories and evaluated accordingly. This section also includes selected examples from the generated surface maps derived from KDE.

The Discussion delineates the contribution of this work, its limitations, and the resulting answers to the initial research questions.

The Conclusion encapsulates the primary findings and provides recommendations in light of future research, concluding with final thoughts.

2 Background

This following chapter on background aims to give a broad overview on the relevant literature and current state of the art relevant for this thesis.

To do this, I will first provide a broad overview of the literature related to naïve geography and the human experience and knowledge of location as and spatial relations. This concept is primarily encapsulated in the notion of naïve geography (Egenhofer 1995), or more specific in vernacular geography.

In the second part, I will expound upon Geographic Information Retrieval (GIR) by elucidating its scope, highlighting both its areas of overlap and distinction from other disciplines, and providing a further description of relevant tasks such as toponym recognition and resolution. Furthermore, this section will delve into the commonly utilized sources pertinent to GIR and central to this thesis, including gazetteers, corpora, user-generated content (UGC), and volunteered geographic information (VGI). Additionally, it will address real estate listings as a subset of VGI.

In the end, there will be a short description on the residential market of Zurich as well, as a background understanding of its dynamics is helpful to understand any possible influence on the data and results.

2.1 Naïve Geography

2.1.1 Historical Contextualization and Concept

This chapter will give a broad overview on the historical roots of the concept of naïve geography and its meaning today. Although the concept of naïve geography has not been defined until as recently as in the 1990s, studies on naïve geography in today's understanding go back a significantly longer way (Egenhofer & Mark, 1995). Even tough not using this specific term, already as early as in the 1960s, Kevin Lynch (Lynch, 1960) conducted experiments on orientation in US cities and grasped some of the aspects that today are summarized under the theory of naïve geography.

Lynch suggests that residents use similar elements to form mental maps for their orientation within a city. By collecting and understanding these individual mental maps Lynch transformed them into a so-called “Image of the City”. This “Image of the City”, that happens also to be the name of his book on the topic, can be built through 5 important elements used by all individuals to create their personal mental map. Lynch concludes from the examined U.S. cities that any other city can also be described with an "Image of the City," which consists of precisely those elements.

Through different qualitative methods such as interviews and sketches with residents, Lynch determined these elements responsible to create such a mental map of the perceived environment. He found out that the following 5 elements to perceive and orient oneself through urban neighborhoods are crucial:

- **Paths**: Paths refer to the various routesthat people use to move within the city, such as streets, sidewalks, and pedestrian pathways. They should be well-defined, continuous, and easily navigable for an effective city design.
- **Edges**: Edges are the boundaries or transitional zones that define and separate different areas or districts within a city. They can be physical features like rivers, highways, or walls, as well as less obvious divisions like changes in land use.
- **Districts**: Districts are recognizable regions or neighborhoods within a city that have a distinct character or identity. They are defined by a combination of physical, social, and functional attributes. Districts often have specific activities, land uses, or cultural significance that make them unique.
- **Nodes**: Nodes are points or prominent landmarks within a city that attract people and serve as points of reference. These can include public squares, intersections, major buildings, or significant gathering places. Nodes play a crucial role in orienting oneself and navigating through the city.
- **Landmarks**: Landmarks are prominent, easily identifiable features or objects that act as reference points in the urban landscape. They can be both natural or human-made, such as monuments, iconic buildings, mountains, or distinctive trees. They are memorable. (Lynch, 1960).

Through the acknowledgement of the individual orientation through mental maps Lynch was among the first to turn his eye to the subject matter that decades later would be captured by the concept of naïve geography. Of course, there are other fields and studies between the time of

1960 to the 1990s that were also concerned with different aspects of naïve geography. Mark and Egenhofer (1995), themselves name for example work in the fields of spatial (information) theory, environmental psychology, artificial intelligence, (naïve) physics, medicine, biology, academic geography, and further.

Mark and Egenhofer (1995), however, were the first to collect these different theories and aspects and create and label a theory on its own to address the topic and its resulting issues for different fields and especially for GIS (Egenhofer & Mark, 1995), which would later serve as an important foundation for further research in the field and related applications of GIS (e.g. (Zelianskaia et al., 2020), (Fogliaroni & Hobel, 2015), (Hollenstein, 2008) and many more.)

According to Egenhofer and Mark (1995), naïve geography refers to individuals' common-sense knowledge and reasoning abilities about spatial relationships and their understanding of the surrounding world. It encompasses very broadly all the informal, intuitive knowledge that individuals possess about spatial relationships based on their experiences and interactions with the environment that are mainly formed by so-called geographic reasoning, that contains a spatial and temporal component (ibd.).

Egenhofer and Mark (1995) declare the theory as the most common and basic form of human intelligence and introduce several concepts and constructs to capture the key aspects of naive geography. One such construct is the "common-sense constraint," which represents the expectations individuals have about spatial relationships based on their everyday experiences. For example, people generally assume that roads follow a continuous path and do not have gaps or abrupt discontinuities. Another concept is the concept of "qualification," which refers to the idea that spatial relationships can have varying degrees of certainty or ambiguity.

It is now visible how naïve geography can be traced back to Lynch as he also addresses the individual’s experience and understanding of space. But, of course, naïve geography is not only concerned with the exact study of Lynch and only complements a name to it. In contrast to Lynch’s focus on the orientation of residents within a city, naïve geography is broader and not bound to an urban area, as it can be applied literally to every place / location / area / etc. that individuals possess spatial knowledge about. Also, the concept is more focused on spatial knowledge of the individuals about certain locations / areas / etc. while Lynch seemed to be more focused on the opposite: He did gather the spatial knowledge of the population, but not with the primary goal of learning more about the specific cities under consideration. Instead, he aimed to abstract a general model ("Image of the City") that is applicable to all cities.

Lynch did, however, also grasp some patterns in his elements that would later be described in the concepts of naïve geography. For example, he acknowledged that the element path should be continuous. This was later generalized by Egenhofers' (1995) concept of "common-sense constraint" which describes general expectations of individuals on spatial relations, for example, a path being continuous. "Common-sense constraint" can be seen as the explanation for why Lynch discovered the importance of continuous paths. The same is true for the element of the district, which is unlike the elements landmark or path not bound to a specific line or point, but a rather vague and possibly also overlapping area. This vagueness can be found in the concept of qualifications, which refers to the idea that spatial relationships can have varying degrees of certainty or ambiguity or even in more depth in the concept of vernacular geography and so-called vague and fuzzy places (see following section 2.1.2 Vernacular Geography).

2.1.2 Vernacular Geography and fuzzy / vague places

Vernacular geography is a more specific field within the rather broad concept of naïve geography (Ettlin, 2011). Unlike naïve geography, vernacular geography is more focused on used terms by individuals in their everyday language. Such terms could refer to a certain area as "downtown" (Evans & Waters, 2007), a spatial relation such as "nearness" (Duckham & Worboys, 2001) that could occur between regions or points (Schockaert et al., 2008). These terms origin from the human experience and knowledge about their environment, why vernacular geography belongs to the whole theory of naïve geography.

The areas are often aligned to a certain human behavior. This is the case because they often include associated information to the environment or socioeconomics and consequently often align with a psychogeographical area. This psychological aspect drives the human activities within these areas (e.g., a "high crime area" that is avoided in the dark or "downtown" where social gatherings take place). (Evans & Waters, 2007).

These vernacular areas do, however, equal to naïve geography (Egenhofer & Mark, 1995) due to their subjective nature lack a clear definition and do consequently distinguish from a formal geographical definition of space (Evans & Waters, 2007). This usually results in not discretely delimited boundaries (ibd), which makes it, on the other hand, clear why they are not suited for "official" geographies that often require clear boundaries to define an area.

According to Evan and Waters (2007) a fuzzy delineation of these vernacular geographies is favored by the following characteristics (a – f):

(a) *Indifference*

Humans often use geographical terms without concern for their precise boundaries. For instance, they decide on a point-by-point basis whether some place belongs to an area or not without knowing an exact expansion of the questioned area ((Evans & Waters, 2007). In this context, Montello (Montello et al., 2003) also refers to a so-called "judgment call" on whether an object belongs to a certain area or not. Schockaert et al. (2008) argue that especially non-political regions (as e.g. the alps) have no need for clear boundaries.

(b) *Continuousness*

Boundaries for vernacular areas often not follow a physical gradient that would make the position of the boundary obvious. In the example of "downtown" it is rather a continuous change from areas with a high density of shops, restaurants, etc. to areas with less density of the same features which makes it difficult to place a boundary (Evans & Waters, 2007).

(c) *Poor precision*

In situations with poor precision, people may use vague boundaries when analytical evidence or inductive suggestions of discrete boundaries exist, but our ability to pinpoint them is limited by measurement techniques or representation methods. (Evans & Waters, 2007)

(d) *Multivariate classification*

When an area is determined through different characteristics with different geographical extension are merged this creates automatically a diffuse boundary (Evans & Waters, 2007).

(e) *Averaging*

Unprecise boundaries can be the result of an average of time or scale-varying boundaries that are associated with one entity (e.g., a hill whose extent varies depending on the scale of observation) (Evans & Waters, 2007).

(f) *Definitional disagreement*

> When areas have varying meanings to different people a definitional disagreement leads to vague boundaries (Evans & Waters, 2007, Davies et al., 2009). Montello (Montello et al., 2003) argues in this context also, that the definition can vary for the same person depending on the context a vernacular area is used (e.g. different extend of "downtown" if one is thinking about the area for shopping, dining or cultural experience).

Especially for vernacular toponyms or regions, the terms "fuzzy places" or "vague places" has been established in various works that try to model or quantify such places. Jones et al. (2008) speak of vague places when they model places using knowledge collected from websites.

Despite the unclear extend and location of these fuzzy places and spatial relations, they are used widely in our communication and still have a specific meaning to the communicators (Duckham & Worboys, 2001) as people not only communicate but also think in vague concepts (Montello et al., 2003). Davies et al (2009) refer in this context to a general inability to even define a place, be it in a spatial or semantical context that makes the concept of place complicated and simple at the same time. This inability of spatial precision is mirrored in the natural language that usually has also not clear semantic boundaries (e.g., "near", "around", "Midwest", etc.), but is nevertheless precise enough to deliver sufficient information to the counterpart (Montello et al., 2003) and the used language and naming is in the end, what space makes to a place and provides it with meaning (Cresswell, 2004).

2.1.3 Relevance of naïve and vernacular geography

The concept of naïve and vernacular geography raises the question, why and in what context it is of relevance. The different available studies claim various reasons which shall be described in the course of this chapter. The focus shall herby lie on geography-related sciences and applications as other fields are of less relevance for this thesis.

The concept is literally everywhere, which means that the knowledge attached to it is also enormous and consequently of interest. As every human's ability to grasp and comprehend space is highly important to get by in private and professional life, mental representations of space are built for almost every aspect of life (Zelianskaia et al., 2020). These mental representations include not only subjective experience but also external sources such as scientific discourses, media, social environment, etc. (ibid) and are therefore a collection and connection of the most diverse sources of knowledge. It therefore makes sense that there is interest in such a profound body of knowledge that is base for all aspects of life.

Of special interest is hereby the knowledge of residents as they generally know most about their environment. Although humans gather and build spatial knowledge for every experienced place and among different scales (e.g., for a very small scale such as a building, or a comparably large scale as a whole city or region, etc.) (Montello, 1998), they include not only present interaction with the environment but also past experiences and memories. This makes the process of spatial learning ongoing over decades during a human's whole life (Montello, 1998, Ishikawa & Montello, 2006). This leads, of course, in general to greater knowledge about the places we spend the most time in, where the knowledge can be even more extensive and detailed (ibd).

There is an interest to understand more about omnipresent human spatial reasoning and decision making (e.g., Egenhofer & Mark, 1995) as it is base and framework for spatial behavior (Montello, 1998) An example for spatial behavior would be navigation or wayfinding; the efficient movement through space to reach one's destination (Montello, 2009). There are different circumstances or elements that support human wayfinding as for example landmarks (Lynch, 1960), a high degree of differentiation, meaning that different parts of the environment do not look alike, or visibility (Montello, 2009). Understanding more about people's conceptualization of space and can thus provide valuable insights for other use, such as the creation of maps, disaster response or urban planning (Hall & Jones, 2022, Hu et al., 2019).

There is also a variety of applications and use for spatial language. In addition to the internal processes that a person goes through, communication with other people also takes place about the environment via spatial language (Hall & Jones, 2022). Spatial language is based on apparently vague spatial relations as "near", "around the corner", etc. (e.g., Egenhofer & Mark, 1995, Montello et al., 2003). If we understand more about this spatial language, it can be used in various applications such as directions in a navigation systems or automatically generated language for descriptions (Hall & Jones, 2022), that among the more recent technical developments are also part of language-based Artificial Intelligence.

Vernacular places can be used to enrich official gazetteers. Most gazetteers origin from the topographic maps created by national mapping agencies and do consequently reflect either official or administrative boundaries or locations (Twaroch et al., 2009). Vernacular places, on the other hand, may differ either in their total existence in the official gazetteer or have the same name, but a different spatial extent (ibd). This human knowledge of space is unlike in the formal world, vague (Schockaert et al., 2008) and stands thus in contrast to all formally used geographies as for example coordinates, or clearly defined points, shapes, and boundaries (Evans & Waters, 2007). This is also one of the reasons these features are often not recorded in official gazetteers, as these vernacular places have no discrete boundaries (Hu et al., 2019).

This understanding leads in more detail and relevance to GIS also to an improved interaction between user and machine (Montello et al., 2003, Twaroch et al., 2009). Usually without any of this information, the interaction between user and machine for any query is very complex and requires several attempts to gain the requested results (Montello et al., 2003). As pointed out by Twaroch et al. (2009) searches often fail as people use place names that are not part of the official gazetteer, why the searching machine cannot fulfill the request. Enriching official gazetteers with local knowledge of vernacular placenames would queries allow to become more efficient (Montello et al., 2003, Twaroch et al., 2009) and allow improved access to geoinformation for lay-people (Hall et al., 2011).

Thanks to its ubiquitous presence and influence in various aspects of life (Zelianskaia et al., 2020), the topic is not only of interest to geographical-related sciences and applications but also to diverse other disciplines (Montello, 1998). This can include work in the fields of psychology (such as behaviorism), language use, neuroscience, computer science, biology, physics, medicine, and further (Montello, 2009, Egenhofer & Mark, 1995).

2.2 GIR

In this chapter, we delve into the more technical and practical background of the topic. While the more abstract concepts of naive/vernacular geography and the significance of place names have been explained in the previous sections, this chapter on Geographic Information Retrieval (GIR) focuses on a more practical use and implementation of this concepts.

Since the relevant disciplines and processes overlap at times, making them not always easy to distinguish, this chapter begins with a schematic overview of the disciplines relevant to GIR and this work (see Figure 2.1). Starting with an introduction to Natural Language Processing (NLP), which is a very broad discipline that (almost) spans the entire topic, a brief overview of Named Entity Recognition (NER) is provided, which overlaps with GIR for the entity "location". In the next section, GIR is introduced as a distinction and extension from Information Retrieval (IR). The final section provides a more detailed description of the particularly important challenges and processes within GIR, such as the process of Geotagging, including Toponym Recognition and Toponym Resolution. Here, the issue of toponym ambiguities, which is one of the biggest difficulties of GIR, is especially relevant.

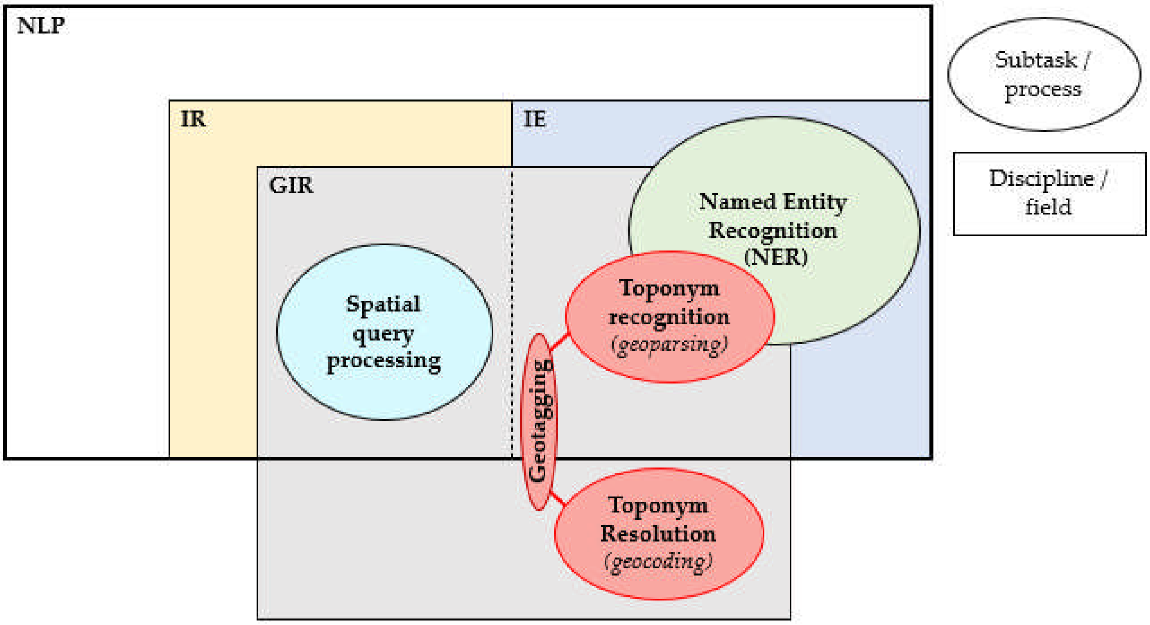

NLP
IR
IE
GIR
Spatial query processing
Named Entity Recognition (NER)
Toponym recognition (geoparsing)
Geotagging
Toponym Resolution (geocoding)
Subtask / process
Discipline / field

Figure 2.1 Schematic overview of GIR-related processes and fields relevant for this thesis

2.2.1 Natural Language Processing (NLP)

As visible in Figure 2.1, NLP is the broadest field. It can be assigned as part of the field of AI, while having received most contributions from the fields of (cognitive) psychology, linguistic, and computer science (Liddy, 2001). Liddy (Liddy, 2001: 1) defines it as following:

> "**Natural Language Processing** is a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks and applications."

She (Liddy, 2001) then explains the different elements in more depth, as they seem on a first glance very broad. But this must be the case in order to grasp all cases with one definition of such a broad field. For example, "naturally occurring text" may sound a bit blurry, but it can literally refer to any text, even oral texts, of any language, mode, genre, etc. – as long as it is natural and not generated by AI (Liddy, 2001). She (ibd.) chose the last part of the definition ("for a range of tasks and application") to emphasize that the field is a very practical one. This is because the main goal of NLP is usually to solve a specific task (e.g., translating a text) or, as in the case of (G)IR, to find specific information, for example, in response to a (spatially relevant) user query to a search engine.

Beside IR and GIR, NLP is concerned with a lot of other subtasks. Further examples would be Information Extraction (IE), question-answering, summarization, machine translation, text generation, and dialogue systems (Liddy, 2001). IE might sound similar to IR, but the two of them have a different focus (Ettlin, 2011; Liddy, 2001).While IR is concerned with the retrieval of relevant documents, IE is concerned with the extraction of the relevant Information within documents (Ettlin, 2011).

2.2.2 IR and GIR

Information Retrieval (IR)

Information Retrieval means according to Collins dictionary "the process of recovering specific information from stored data" (Collins English Dictionary, 2023) or alternatively, according to Cambridge Dictionary "the process of finding stored information on a computer" (Cambridge Dictionary, 2023). Information Retrieval (IR) as a scientific field, however, is not as broad as the common understanding of the term (Manning et al., 2009). Within the scientific context the finding of this information is a) within a large amount of data (where traditional cataloguing techniques reach their limits) (Sanderson & Croft, 2012) and b) the searched data must be not structured or at least only semi-structured, which applies in most in cases to text information (Manning et al., 2009; Sanderson & Croft, 2012).

The field of study of Information Retrieval (IR) is quite old, dating back to Calvin Mooers in 1950 (Mooers, 1950), who first introduced the term (Swanson, 1998). The field of IR emerged as a response to the problem of a large amount of data, wherein a user needs to access specific information, and how, to reach this goal, scientific articles and reports should be indexed (Swanson, 1998). Although the problem may not have been addressed under the term "information retrieval," it arose even earlier, such as during the management of extensive collections of information in libraries (Sanderson & Croft, 2012), or an even earlier example would be the first creation of an index to the bible by monks in 1247 (Swanson, 1998).

Although in the beginnings of IR, the data addressed was mainly scientific articles and reports, today the information is not only limited to these two categories but can be anything if it is unstructured or semi structured data (Sanderson & Croft, 2012) and consequently non-numeric (Mooers, 1950). Examples for such unstructured or semi structured data would be web pages, documents, images, or videos (Sanderson & Croft, 2012), while text information is the most often targeted information (Manning et al., 2009). Especially with the rise of the world wide web, these data sources have enlarged enormously (Sanderson & Croft, 2012).

The world wide web and rise of information technology did add further possibilities to IR. As while the field and the problem it addresses have been known for a long time, IR systems, the resulting technology, have significantly grown in their capabilities with the advancement of computer technologies that provide greater processor speed and storage capacities and large sources of information through the world wide web (Sanderson & Croft, 2012).

However, not only have the capabilities of IR expanded with the internet and technological progress, but its meaning and significance for people today is also much bigger. At the outset, it was mainly specialized professions that came into contact with an IR system, as for example in the mentioned libraries. Today, anyone who performs a data query through a search engine like Google is already in contact with an IR system (Manning et al., 2009).

Geographical Information Retrieval (GIR)

Geographical Information Retrieval is not only a part but rather an extension of IR (Jones & Purves, 2008) and does, as the name already reveals, add a specifically geographic focus on the retrieved information. Unlike to IR, the separate field of GIR is rather new and has emerged around the turn of millennium, why it is probably said to have a focus on documents gathered from the web in the first place (Jones & Purves, 2008). GIR can consequently be defined as a discipline focused on creating search systems that are sensitive to spatial information and assisting users in fulfilling their geographic information requirements (Purves et al., 2018).

GIR as an emerging separate field is of relevance because the geographic context is in itself of great relevance (see also section 2.1.3). This is the case as a lot of searching requests for information retrieval have a spatial relevance. According to (Jones & Purves, 2008) approximately 13-15% of web queries on search engines contain a place name. This makes sense, as a lot of the information a user needs has is directly connected to its own location. Common examples for such requests are searching for opening hours and/or location of nearby businesses, or local product availability. In this context, the advertisements displayed to the user are often related to the user's geographical location as this is most effective. (Purves et al., 2018).

Equal to IR, the searched data is unstructured as well (Jones & Purves, 2008), but the link to the geographic information can have different forms. On the one hand, it can involve direct links to a location when coordinates are available, as is the case with "geotagging" (Hu, 2018). An example of this would be the social network Flickr, where users tag and upload pictures with geocoordinates (Hollenstein, 2008) or geotagged Wikipedia pages or tweets (Hu, 2018). In the much more common scenario, the link to a location, however, is not geocoded but is only part of unstructured text or so-called natural language containing a location description. This is called geo-text data and can be found in e.g., news articles, travel blogs, historical archives, etc. as long as there exists a connection between location and text (Hu, 2018).

While the detection of the geographic references can be challenging on its own, the field of GIR is also faced with further challenges. The most important ones have been summarized by Jones and Purves (2008) and are namely the following 7 issues:

1) *Detecting geographic references*

   Through the process of toponym recognition (Lieberman et al., 2010), as during a Named Entity Recognition (NER) (see section 2.2.3), the geographic information (place name) can be extracted out of the unstructured (text) information. This process deals with the two main challenges of place names that are not part of official gazetteers and have thereby to be detected through these linguistic approaches (Twaroch et al., 2009), or, there place names mentioned that in this specific context do not refer to a placename because they are used metonymically (Jones & Purves, 2008).

2) *disambiguating place names* (see also 2.2.4)

   After the toponym recognition, the found toponyms must be assigned the correct geographical coordinates in a second step. This step of assigning to the correct location is called toponym resolution. The entire process of toponym recognition and resolution is then summarized under the term geotagging. (Lieberman et al., 2010).

During the process of toponym resolution, the issue of toponym ambiguities becomes particularly relevant (Jones & Purves, 2008; Lieberman et al., 2010). These ambiguities occur as some toponyms can be interpreted differently, as it is the case when e.g., the name of a city exists multiple times for different cities (Lieberman et al., 2010).

3) *vague geographic terminology*

As already described in the section on naïve geography (see chapter 2.1), people think, resonate and communicate in vague spatial concepts with often no sharp boundaries (Montello et al., 2003). Consequently, there is a need for solutions to grasp the vagueness of these vernacular places or spatial relations, which is often difficult as official geographies work with sharp defined boundaries, distances and coordinates (Evans & Waters, 2007; Jones & Purves, 2008).

4) *spatial and textual indexing*

When categorizing web documents based on their geographical context, they need to be indexed for quick retrieval in response to user queries (Jones & Purves, 2008). This also the case for IR, where e.g. any searching machine needs to be able to rank the results according to their relevance for the user's request (Manning et al., 2009). According to Jones & Purves (2008), for GIR, this task involves creating an inverted file associating words with documents and combining it with a spatial index for geographic relevance. Document footprints represent geographic references in documents and are used for indexing. The challenge hereby lies in efficiently combining text and spatial indexes. GIR queries are even characterized as triplets as they consist of a topic, spatial relationship, and location. (ibd).

5) *geographical relevance making*

After identifying potentially relevant documents, the next challenge is ranking them based on relevance to the query. The relevance focusses on the two factors of the frequency of query terms in retrieved documents and the geometric match between the query and document footprints. These two scores can then be combined to calculate an overall relevance score. (Jones & Purves, 2008).

6) *user interfaces*

GIR poses challenges for user interfaces in query formulation and result presentation. GIR queries are often triplets as they combine a topic, spatial relationship and location, which can be addressed through structured interfaces or map-based approaches. Map-based visualization is common, but challenges include aggregating documents with the same footprint, handling documents with extensive geographic scopes, and utilizing geovisualization techniques to explore large document sets returned by GIR systems. (Jones & Purves, 2008).

7) *user studies and evaluation*

As in the end, any GIR system must fulfill user's requests, it is of great importance to include those user needs into the development of the system. But to gain insight into existing gaps and analyze where requests fail or need multiple trials, is very challenging as unfortunately there is not much data existing on query logs that could be used for this research (Jones & Purves, 2008).

2.2.3 Named Entity Recognition (NER)

As visible in Figure 2.1 the task of NER, is a classical task of IR (Goyal et al., 2018) and has, when it comes to the recognition of the entity of location, a shared part with GIR. A NER is not only part of classical IE but can also work as a crucial foundation for many further fields or tasks that would typically be assigned to NLP, such as IR, question answering, machine translation, automatic text summarization, text clustering, knowledge or ontology-based population, opinion mining, semantic search, and many others (Goyal et al., 2018).

So far, a lot of foundational work for NER has been made due to the Message Understanding Conferences (MUC) (Chieu & Ng, 2002). The MUCs originated in the USA and primarily served to stimulate research in the automatic analysis of military messages (Grishman & Sundheim, 1996). In this sense, they are not a conference in the traditional sense, because in order to participate, a research team had to create an IE system and compete against other teams. The generated systems were then tested by releasing a previously unknown test dataset just before the conference, which the participants then fed into their system (Grishman & Sundheim, 1996). While the first MUCs were concerned with classical IE, the MUC-6 and MUC-7 then focused on

NER (Chieu & Ng, 2002). In these two conferences, the core aspects of a Named Entity Recognizer (NER) emerged.

As the name already reveals, NER is concerned with the recognition of named entities. According to Goya et al. (2018 : 22) a named entity can be defined as follows: "A named entity is a word form that recognizes the elements having similar properties from a collection of elements." An example for such an entity type would be persons or locations (see Table 2.a). Chinchor (1997) defined the first applied entity classes during an MUC. She suggests (ibd.) the named entity tasks can be divided into a total of 3 subtask, that would be the recognition of entity names, temporal expressions and number expressions that each are divided into different types (see following Table 2.a):

Subtask	**Type**	**Description**
Named Entities	Organization	Named corporate, governmental, or other organizational entity
	Person	Named person or family
	Location	Name of politically or geographically defined location (cities, provinces, countries, international regions, water bodies, etc.)
Temporal Expressions	Date	Complete or partial date expression
	Time	Complete or partial expression of time of day
Number Expressions	Money	Monetary expression
	Perrcent	percentage

Table 2.a Overview on subtasks and entities of NER in MUC-7 (Chinchor, 1997)

The preceding table by no means represents the only classification that would be common in an NER. Instead, it is more of a classical categorization, which could vary in a specific area of interest (Goyal et al., 2018). Goyal et al (2018) mentions biomedicine as an example, where a typical NER would, for instance, detect entities like genes and gene products.

The goal of the NER is then, to mine the text base and deliver the relevant tokens with the right tag (words or word sequences) (Chinchor, 1997). In assessing which system of the competing teams was the best, the evaluation metrics Precision and Recall were invented during an MUC, which still happen to serve as important benchmarks to this day (Grishman & Sundheim, 1996).

The used techniques to build a NER differ from system to system. Two classical approaches are rule-based NER and machine-learning-based NER (Grover et al., 2008). As the name implies, the machine-learning-based NER requires more annotated training data, which typically makes it more labor-intensive than a rule-based NER. On the other hand, a rule-based NER is less adaptive to unexpected entities, as it works with predefined rules. (Grover et al., 2008). A further, even simpler type of NER, would be a dictionary-based approach, where a predefined list of entities is used as a look-up (Jones et al., 2001). This has of course the disadvantage, that everything outside of the gazetteer is missed, as it would be the case for a lot of vernacular placenames (Twaroch et al., 2009).

Beside the used approach, also the language can have an enormous influence on the result as the typological properties of the investigated language entail various difficulties (Baumann, 2019). Most studies for NERs exist for the English language, followed by other European languages. German often performed worse in NER systems than English due to various reasons such as German being more morphologically complex, less performed studies as there is a greater interest in English (Klimek et al., 2017). A central aspect seems to be capital letters, that can work as a great identifier for English or other European languages, but does not work to well for German, where every noun starts with a capital letter, making it difficult to work as an indicator for entity assignment (Goyal et al., 2018).

2.2.4 Toponym recognition and resolution and Toponym Disambiguation

A toponym is a more formal term for a place name. The detection of toponyms is not the only but certainly a key task of GIR (see Figure 2.1) that is often referred to with *toponym recognition* or alternatively *geoparsing* (Purves et al., 2018). The aim of geoparsing is the identification of toponyms out of an unstructured text. These detected place names can be words or even whole phrases (Hu, 2018). In a NER, the detection of the entity "Location" can be referred to as Toponym Recognition (Lieberman & Hanan, 2011). However, since Toponym Recognition can be performed not only through NER (Lieberman et al., 2010), it is not a subfield of NER, even though both belong to IR and share many similarities (see Figure 2.1).

The process of toponym recognition might sound trivial because this is the case for us humans (Ardanuy & Sporleder, 2017), but it is not for any computer as it involves the resolution of semantic ambiguities (Lieberman et al., 2010). As toponyms or other geographical references are usually highly under-specified and ambiguous (Rauch et al., 2003) and an understanding of the natural language is required (Lieberman et al., 2010). For human beings, however, these difficulties do not exist because they have other abilities to distinguish place names from other names and resolve ambiguities based on their experiences and real-world knowledge (Rauch et al., 2003).

There are several examples one could provide for a better understanding what a semantic ambiguity means. A first good example would be the name *Jordan* (Lieberman & Hanan, 2011). On the one hand, it could refer to a river or a country, on the other hand, to a person. Another example for a semantic ambiguity would be the metonymical use of a place name as for example in "talks with Washington" as a reference to the government and not the location (Jones & Purves, 2008).

The resolution of these semantic toponym ambiguities can be addressed through different approaches. On the one hand, an additional source can be consulted: According to Ardanuy and Sporleder (2017), Wikipedia is a common tool in this regard, as it has the significant advantage of describing entities in a context provided by natural language. Other approaches are more focused on the language itself, as e.g. in the study of Lieberman and Samet (Lieberman et al., 2010), where different syntactical and grammatical approaches were applied (e.g. position of the toponym in the sentence, only passive verbs in connection with a toponym, etc.).

In order to receive a valuable result, the process of toponym recognition, is often enriched with the second step of *toponym resolution*, also referred to as *geocoding* (Lieberman et al., 2010). The combination of geoparsing and geocoding, is also referred to as *geotagging* (ibd.), that as a whole, can be regarded as the process to enable spatial indexing to an unstructured text (Lieberman & Hanan, 2011). While toponym recognition deals with the detection of geographical references in general, toponym resolution is the part, where the detected geographical reference is assigned to a place on earth, respectively quantified with coordinates (Lieberman et al., 2010; Lieberman & Hanan, 2011).

Toponym resolution has to deal with ambiguities as well, but unlike as in toponym recognition they are not of a semantic nature but of a spatial one (Lieberman et al., 2010). A popular example would be *Madison*, that once detected as a geographical reference and not a family name, still must be assigned to the right city among the other over 24 existing cities in the USA with the same exact name (Rauch et al., 2003). This form of ambiguity is further compounded when the place name does not necessarily refer to different places, but to distinct geographical concepts (Batista et al., 2012). Batista et al. (Batista et al., 2012) bring up the example of Lisbon, which can have multiple referents as street names, community, city, or region, some of which may also coexist within the same geographic area. Spatial ambiguities become even more complex when dealing with a dataset of historical locations, as their meaning and names can change over time (Ardanuy & Sporleder, 2017). Also, the ambiguities differ for the different entity types:

In order to resolve these ambiguities and determine the right location, one must possess a solid comprehension of the document's content in order to accurately determine which among the numerous potential locations is being alluded to (Lieberman et al., 2010). According to Lieberman et al. (2010) the two pre-dominantly used models do either assume it is always the most referred to location (e.g. Paris, France instead of Paris, Texas) or check for other location references within the same document, assuming the ambiguity can be resolved by picking the location nearer to the other mentioned places names. The second approach is more sophisticated and seems to underline Tobler's 1st law of Geography: "Everything is related to everything else, but near things are more related than distant things" (Tobler, 1970 : 236). Brunner and Purves (2008) have come to a am similar conclusion, when showing in their study that ambiguous toponyms are spatially autocorrelated.

According to Buscaldi (2011) the approaches to address toponym ambiguities can be divided into three main categories: map-based methods, knowledge-based methods, and data-driven or supervised methods:

1) For **map-based methods** coordinates have to be already existing (see section 2.3 for examples where this would be the case). Then different spatial approaches can be used such as calculating the average distance between toponyms and remove "false" toponyms that are too far away (Buscaldi, 2011). Other methods include the modelling of spatial extent through density surfaces (Jones & Purves, 2008). These methods are especially useful for vernacular geographies with no sharp boundaries (ibd).

2) **Knowledge-based approaches** include external knowledge sources such as gazetteers (see section 2.3) or ontologies (Buscaldi, 2011). An example would be the use of inhabitant numbers, assuming that more densely populated areas are more likely to be mentioned or Wikipedia as a source to include templates, categories and referents for a further clue (Buscaldi, 2011).

3) **Data driven or supervised methods** base on machine learning techniques and are not often used due to a lack of geotagged data and unseen toponyms that cannot be

classified (Buscaldi, 2011). However, when used in in combination with non-geographical data to build a model based on the spatial relations between this non-geographical data and the toponyms, the data driven method has also led to good results (Buscaldi, 2011).

However, no matter which methods is used, toponym ambiguities remain a big challenge in both, toponym recognition and toponym resolution, and still require more research and approaches to deal with them (Jones & Purves, 2008).

2.3 Sources of GIR

At the conclusion of the section on GIR, this chapter will provide a brief selection of the commonly used sources from which geographic information has been extracted. This completes the overview of GIR and ultimately serves to define the research gap, specifically in terms of which sources, methods, and languages have been less explored thus far. It should be noted that there can also be overlaps between the different type of sources. For example, a corpus can be built out of volunteered geographic data (VGI) (e.g. Wallgrün et al., 2014) or VGI is suggested to be used to enrich a gazetteer (Kessler et al., 2009), resulting in an overlap between the three described sources.

2.3.1 Gazetteers

Gazetteers can be described as toponym dictionaries and are a form of knowledge information system (Hill, 2006), that can be characterized according to their scope, coverage and detail or granularity (Buscaldi, 2011). Digital gazetteers are typically structured as triples, consisting of a place name, geographic footprint, and feature type (Kessler et al., 2009). Since they include functions for mapping place names to their footprint or assigning them to a feature type, they serve a purpose beyond that of a dictionary and constitute the central subset for web feature services that handle place names (Kessler et al., 2009).

While gazetteers are now mostly digital, their origins can be traced back to reference books that recorded place names and their official spellings (M. F. Goodchild & Hill, 2008). Therefore, it is not surprising that the majority of gazetteers is generated by national mapping agencies and does consequently reflect either official or administrative boundaries (Twaroch et al., 2009) and has due to their origin both lead to and is based on administrative names of places (M. F.

Goodchild & Hill, 2008; Hill, 2006), which consequently results in the frequent exclusion of knowledge encompassed within vernacular geography (Jones & Purves, 2008).

Due to their ease of use, gazetteers have already been used for initial attempts to assign a user's spatial search query to a geocoordinate, which in turn serves as the basis for a coordinate-based request within a Geographical Information System (GIS). In this rather straightforward approach, relationships between place names were not yet considered, as it is the case with modern gazetteers. Such in detail gazetteers enable the encoding of semantic or spatial relationships between locations and offer greater flexibility in terms of spatial extent, which is not anymore limited to a point (geocoordinate). They would also include further information such as alternative names for the same locality. (Jones et al., 2001).

Hence, it seems evident that the scope of the chosen gazetteer significantly influences the outcome of an analysis, a search request, or other applications (Buscaldi, 2011) and it is crucial to improve existing gazetteers and create new ones (Goodchild & Hill, 2008; Twaroch et al., 2009). This is reflected in the construction of (digital) gazetteers as a growing scientific field that brings together various disciplines such as geographic information science, geographical information retrieval (GIR), computer science, and social history (Goodchild & Hill, 2008). Especially when it comes to vernacular geography, due to their official nature standard gazetteers usually have certain limitations (Jones & Purves, 2008) that should be improved by adding this knowledge of naïve geography (Egenhofer & Mark, 1995).

Once created, a gazetteer's utility is manifold. E.g., in the process of geotagging, gazetteers are mainly used during toponym resolution (geocoding) as a reference source to assign the recognized toponym to geo-coordinates (Ardanuy & Sporleder, 2017) or they are used as a "look-up" list to mine the unstructured text during toponym recognition (Lieberman & Hanan, 2011). Also, they own a crucial role in GIS and related applications such as web-based mapping services and search engines (Kessler et al., 2009).

2.3.2 Corpora

While gazetteers play an important supporting role in enabling the tasks of toponym resolution and recognition (see 2.3.1), in a GIR context corpora are a crucial source to extract the place information itself. Since toponym recognition focuses on extracting place names from unstructured text, in this case so-called geo-text data (Hu, 2018), it is not surprising that a corpus is defined as a large, systematic digital collection of text, already prepared for a linguistic task (Ettlin, 2011).

These text collections are increasingly sourced from the web and can originate from a variety of sources (Lieberman & Hanan, 2011). Well-known established corpora include collections of newspaper and journal articles, or they are built from volunteered geographic data (see section 2.3.3) or user-generated implicit geographic data such as travel blogs, reviews, and more. This chapter on corpora aims to make it clearer through several examples what is meant in practice by a corpus and what (spatial) information could be extracted from it.

Well-known and extensively studied corpora are those containing collections of newspaper or similar press articles because reporting typically involves many spatial references (Lieberman & Hanan, 2011; Pasley et al., 2008). They can be characterized as rather formal in comparison to other corpora built out of co-authored and less formal content (Pasley et al., 2008). Teitler et al. (2008) underline the need to examine news corpora by pointing out that in many online news channels, the geographical references are not prominently identifiable. They tend to group articles based on topics rather than geographies, even though the latter are equally if not more of interest (2008). Similar characteristics (formal, often include geographic information) can be found for a corpus consisting of scientific articles, as for example used in the study of Acheson and Purves, where they processed a toponym recognition on a corpus of scientific articles (Acheson & Purves, 2021).

Another example, perhaps somewhat specific but intriguing for this thesis due to its geographical context (Switzerland), is the Text+Berg Corpus. It consists of the digitized Swiss Alpine Club Year books back until 1864 (Derungs & Purves, 2014). Derungs and Purves (2014) georeference the toponyms contained in the over 10,000 articles, map them, and identify selected natural features within them. These features are then placed into a spatial and temporal context to measure individual regions in Switzerland and the similarities between them. Ettlin (2011) used the very same corpus in her thesis to extract vernacular and alternatively spelled toponyms, considering that these could be needed for the Swiss Air-Rescue Service (REGA) to quickly locate an accident site in the mountains.

Less formal corpora consist of user generated content (UGC), that is created by individuals and combines various forms such as blogs, tweets, or other sources of social media (Hollenstein & Purves, 2010). The prerequisite of UGC is that these are data uploaded by individuals on the web (Hollenstein & Purves, 2010). Since UGC working for a corpus in GIR require a spatial reference, volunteered geographic data (VGI) is typically used as a specific subset of UGC (Goodchild, 2007). VGI, as an increasingly important source (Goodchild, 2007), is discussed separately in the following chapter (see section 2.3.3).

However, many studies in GIR do not only deal with the extraction of toponyms or spatial relation expressions (SRE) from a corpus but also focus on the task of utilizing them as a basis to create a new corpus. For example, Chesnokova and Purves (2018) have created a georeferenced corpus based on extracted first-person perceptions of landscapes in the English Lake District, primarily collected by web crawling from different sources such as travel blogs. Thus, the collected descriptions of landscape perceptions also constitute a corpus. Two others of numerous examples are the study by Wallgrün et al. (2014), who generated a corpus for Spatial Relationship Expressions (SRE) from a corpus of hotel reviews and the study of Gelernter and Mushegian (2011), who build a corpus based on tweets mentioning harmed locations during the earthquake in Christchurch, New Zealand in 2011.

2.3.3 Volunteered Geographic Information (VGI)

The last described important source is known as Volunteered Geographic Information (VGI) which has developed in conjunction with the web. The ever-expanding web holds vast amounts of information that can be analyzed across various disciplines (Pasley et al., 2008; Sanderson & Croft, 2012). However, as the data found on the web often has a spatial reference, making it the largest collection of spatial data, the web is particularly intriguing for GIR (Pasley et al., 2008). A significant group of such used spatial data in GIR, which has grown alongside the web and is increasingly being considered, is VGI (Pasley et al., 2008).

VGI can be considered as a special case of user generated content (UGC) (Goodchild, 2007) and does, unlike UGC, necessarily include a spatial reference (Hollenstein & Purves, 2010), which makes it especially interesting to gain insights into vernacular geographies (Evans & Waters, 2007). This spatial reference can either be added as unambiguous coordinates or less clearly apparent, as toponyms that exist in various forms, such as part of a description or as tags, as they are often found in social media (Hollenstein & Purves, 2010). Since the creators of such VGI are both, users and producers the term 'produsers' has emerged to describe them (Coleman et al., 2009).

The most famous example for VGI is OpenStreetMap[1] (OSM), a crowdsourced database, that is often used to exemplify VGI (Mooney & Minghini, 2017). OSM was founded in 2004 with the idea that local volunteers, who possess local knowledge of their respective areas, would collect geographic data. The combination of all these contributions would ultimately lead to a crowdsourced map for the entire region, or ideally, the whole world. (Mooney & Minghini, 2017). Today, the core of OSM is a spatial database that contains geographic data and information from around the world (Mooney & Minghini, 2017).

In other sources, there are no deliberately indexed projects serving the collection of VGI; rather, these are more of a byproduct, as is the case with certain social media platforms. An example is the image database Flickr, where users post georeferenced images with descriptions. Hollenstein and Purves used these to determine the locations of vernacular downtown areas in Zurich City and other cities (Hollenstein & Purves, 2010) and Bahredar et al. used the meta-data attached to Flickr images (and OSM) to create an interactive image of the city of London, grasping the perception of citizens on certain city areas (Bahrehdar et al., 2020).

Beside Flicker, there are also other examples of social media in VGI: Amon many other a very popularly used one are georeferenced microblog entries in Twitter[2] (Bahrehdar et al., 2020). One particularly intriguing application of VGI data from Twitter is disaster management, for which Twitter data from all social media have proven to be the most helpful (Granell & Ostermann, 2016). As social media plays an important role to communicate during large-scale disasters and can contain different relevant pieces of information, as e.g. extent of damage or medically related help calls (Zahra et al., 2022).

Probably one of the biggest challenges related to VGI is the data quality due to the potential for contributions from anyone (Goodchild & Li, 2012). There are different studies addressing this issue by investigating on different factors that determine the character of the VGI such as the role of the contributor (e.g. its motivation) (Granell & Ostermann, 2016), and thus different approaches to assess quality (M. Goodchild & Li, 2012).

2.3.4 Real estate data

According to Hu et al. (2019), real estate data (listings) also fall under the VGI data type, as introduced by Goodchild (Chen & Biljecki, 2022). Since the present thesis is based on a dataset

[1] www.openstreetmap.org

[2] www.twitter.com, although url contains Twitter, the name has been recently changed to 'X'

of listings, special attention is given to this data category, and it is treated separately from the chapter on VGI.

While real estate data has not been as frequently used in the domains of GIR to date (Chen & Biljecki, 2022), it has some significant advantages compared to other sources of VGI, especially social media (Hu et al., 2019):

1. Since the location of a housing unit is its most important selling factor, it is described comprehensively and thus, mining of vernacular geographies in listings is especially promising.
2. The same is true for nearby features, that are described a lot and thus real estate listings hold potential to gather information about them, or about SRE to them.
3. Listings do usually have less noise and more geographic connection to a place as in social media, where the user can post from anywhere.
4. Listings hold potential to show better coverage over the whole area (city), as social media often targets more touristic areas.

These advantages are utilized by various studies:

Hu et al (2019) themselves extracted place names from advertisements in different cities and cartographically mapped them as spatial footprints based on the pre-existing geocoordinates (address of the advertisement). Subsequently, a comparison could be made with official gazetteers, thereby detecting previously unrecorded local place names.

Chen and Biljecki (2022) crawled listings from the Singaporean real estate market over a period of three months, focusing on the extraction of sports facilities from the advertisement texts and images. They geocoded the addresses found in the advertisements, allowing them to assign a location to each facility. This location data was then cross-referenced with existing data in OpenStreetMap (OSM). In doing so, they developed a method for assessing the quality of an existing database (OSM) and potentially expanding it using extracted information from the advertisements.

Nilsson and Delmelle (2023) conducted a study in Mecklenburg County, North Carolina, wherein they employed text classification techniques to investigate micro-geographic housing dynamics. This involved the categorization of real estate listings based on property life cycles using predefined classes and keywords. Subsequently, they spatially visualized these classified data points across various time intervals on a grid representation, facilitating a nuanced examination of temporal changes within the local real estate market.

Brunila et al. (2023) center their study on the advertisement descriptions within the short-term rental market, exemplified by platforms such as Airbnb. In doing so, they scrutinize these descriptions for place names and spatial relationships with the aim of quantifying these elements and subsequently comparing them with reviews authored by guests of the short-term accommodations. Their findings primarily revolve around the concept of 'proximity,' revealing disparities between the way proximity is presented in listings compared to how it is perceived in the reviews.

The strong relationship between real estate and location is also leveraged by Wallgrün et al (2014) in their study, based on hotel reviews. Although these are not real estate data in the form of listings as in the previously mentioned studies, the description of the location is still an important component described in the examined reviews. These hotel reviews are now being incorporated to build a corpus for natural language expressions, extracting toponyms along with spatial expressions such as "near" or "to the left of" (Wallgrün et al., 2014) This is intended to assist in better understanding Spatial Relationship Extraction (SRE), which, in turn, is crucial for a wide range of applications, including, among others, IR and human-machine interactions (Wallgrün et al., 2014).

2.4 Residential market of Zurich-City

Given that this thesis is working with a dataset of real estate advertisements from Zurich-City, it is consequently significantly influenced by the residential market as a whole. Therefore, this chapter aims to provide a comprehensive overview of the residential market in Zurich-City. It will describe the historical development of the market in relation to and in differentiation from other parts of the country, both in terms of its supply and demand, as well as the thereof resulting consequences and dynamics.

2.4.1 Structure of Zurich-City

With a population of 426,890 inhabitants as of 2022 (Kanton Zürich, 2023), Zurich-City is the largest city in Switzerland. The residents are distributed over 12 city districts (in the local language called “Kreis”). All districts can be referred to by their number (e.g. “Kreis 4”), but some of them also have alternative name (e.g. "Kreis 5" or "Industriequartier"). The districts in turn can be further subdivided into 1 to 4 city quarters. In total there are 34 city quarters (Zürich Tourismus, 2023). The name of the district or quarter can sometimes refer to a name of a former community that has been integrated in the city (e.g. "Wiedikon" (district 3) or "Altstetten" (quarter of district 9)).

The residential population is not evenly distributed across the various city districts (see Table 2.b): In absolute numbers, the majority of the population resides in District 11 (17.6%), followed by District 9 (13.2%). The district with by far the fewest residents is District 1 (1.3% of the population), which houses the historic old town and has more commercial than residential areas. However, the districts vary greatly in size, so it makes sense to consider population density as an additional indicator. Here, the picture looks slightly different; the most densely populated districts are also centrally located, Districts 3 and 4, which have population densities of 145 people per hectare and 193.8 people per hectare, respectively. While all districts have shown an increase in population over the past two decades (year 2000 to year 2022), by far, District 11 has experienced the most growth in recent years, with an increase of 44.2%. The inner, already densely built-up areas could comparatively only record lower growth rates.

District	**Population (2022)**	**Share on total Population in %**	**Development of Population (2000 – 2022)**	**Population density (Pers / ha) for urban area only**
1	5,860	1.3%	+1.3%	67.4
2	37,049	8.4%	+28.0%	104.1
3	50,582	11.4%	+10.8%	145.0
4	29,672	6.7%	+9.8%	193.8
5	15,888	3.6%	+35.2%	123.9
6	35,862	8.1%	+20.4%	134.6
7	39,301	8.9%	+16.6%	80.9
8	17,814	4.0%	+16.1%	96.1
9	58,517	13.2%	+31.1%	109.8
10	41,505	9.4%	+14.9%	117.2
11	78,034	17.6%	+44.2%	124.5
12	32,953	7.4%	+17.3%	121.1
Total	**443,037**	**100.0%**	**+ 22.7%**	**116.7**

Table 2.b Population today, development over past years, and density among different districts in Zurich-City. Table based on (Stadt Zürich Statistik, 2023a, 2023b)

If the same analysis is conducted at the neighborhood level, the strongest growth of +222.7% occurred in the Escher Wyss neighborhood (District 5) (Stadt Zürich Statistik, 2023b). This is not surprising, as there has been significant development in recent years, with former industrial buildings being converted into residential and other uses building now a very trendy city quarter (Stadt Zürich Statistik, 2023c).

2.4.2 Supply

The housing market in Zurich-City has been marked by extreme scarcity for years. As of 2022, Zurich-City has a rental vacancy rate of 0.07%, which corresponds to a total of 161 units (Federal Statistical Office, 2022). Experts believe that any vacancy rate below 1% already indicates a housing shortage (Thalmann P, 2012). Any residential market in Switzerland would be in equilibrium with a vacancy rate of about 1.0% to 1.5% (Thalmann P, 2012), which is the case for the whole country where the overall vacancy rate refers to 1.31% (Federal Statistical Office, 2022). With a vacancy rate of 0.07% so significantly below 1%, the residential market in Zurich-City is extremely dry.

Many other Swiss major cities also exhibit low vacancy rates (see Table 2.c), but the rate in Zurich-City is particularly low. Geneva, Bern, and Lausanne are also experiencing low vacancy rates, while the situation in Basel and Lausanne is notably more relaxed with vacancy rates over 1%. In contrast to the tight supply in cities, some rural communities are struggling with high vacancy rates of up to more than 10% (Federal Statistical Office, 2022), while in the suburbs around the cities, there is often also limited supply due to low vacancy rates as they absorb the increased demand from the city. This low vacancy rates reflect a strong trend towards urbanization.

Largest Swiss cities **(sorted in descending order by population)**	**Vacancy rate** **(as of June 1st, 2022)**
Zurich	0.07
Geneva	0.47
Basel	1.20
Lausanne	1.29

Largest Swiss cities **(sorted in descending order by population)**	**Vacancy rate** **(as of June 1st, 2022)**
Bern	0.25
Winterthur	0.37
Lucerne	0.88

Table 2.c Vacancy rates in biggest Swiss Cities (Federal Statistical Office, 2022)

Zurich-City does not generally have many high-rise buildings. Especially in the residential market, there are currently only 18 residential blocks that reach a height of 55 meters or more (see Table 2.d). Noteworthy are the years of their construction that seems to be dividable into two waves of construction. The first wave of high-rise buildings occurred in the early late 1950ies to 1970s (e.g. Lochergut, Hardau Towers). This initial wave was followed by decades of hiatus. Since 2010, residential high-rises have been constructed again (e.g., Vulcano Towers in Altstetten) or are in the planning stages. Among the planned buildings are the future tallest structures in the city, which will be residential towers. These are two residential high-rises that are intended to be built along with a new football stadium, each reaching a total height of 137 meters, making them the tallest buildings in the city (Amt für Städtebau & Stadt Zürich, 2022; Von Ledebur, 2023). However, with the recent decline of Credit Suisse Bank, which is the owner, the current status of the project is currently uncertain (Von Ledebur, 2023).

	Construction Year	**Building name or address**	**Height**
1st wave	1959	Schwesternsilo	56m
	1966	Lochergut	62m
	1967	Im Holzerhurd 11	57m
	1973	Schwandenholzstr. 24	57m
	1976	Hardau II	4 towers up to 95.4m
	1976	Sihlweid	2 towers with 58/66m
	1976	GZ Grünau (63 m)	63m
2nd wave	2011	Mobimo tower	81m
	2011	Leutschentower	60m
	2013	Löwenbräuareal	70m
	2013	Vulkanplatz 21-27	60m
	2014	Zölly Hochhaus	76m
	2014	Escher-Terrassen	60m

	Construction Year	Building name or address	Height
	2015/16	The Metropolitans West / Ost	2 towers à 62m
	2018	Vulcano towers	3 towers à 80m
	2018	Labitzke-Areal	2 towers à 46/64m
	2015/2019	Europaallee	Includes 2 buildings of 56 / 60m
	2020	Wolkenwerk	4 towers à 62m

Table 2.d Residential buildings in Zurich-City with minimum height of 55m, sorted by construction year. Table by Chantal Meier based on Amt für Städtebau & Stadt Zürich, 2022

2.4.3 Demand

The waves of construction of high-rise buildings correlate with the city’s population development (see Figure 2.2). In terms of inhabitants, Zurich reached its first peak in the 1960s when the population stood at 437,273, which is almost equal to today’s population (Stadt Zürich Statistik, 2023b). After that the 60ies, there was – as in almost any other Swiss city – a period of suburbanization of so-called "urban flight" during which the population significantly decreased to level of below 360,000 (Statistik Stadt Zürich, 2005). This “urban flight” started with families that left the city during the late 60ies and 70ies, followed by an ever stronger decrease of urban population in the 70ies and 80ies caused by increased crowding costs in the city (pollution, rents, traffic, criminality) and at the same time better public transport connections to the more rural areas (Eichler et al., 2013). However, since the late 1990s, there has been a clear trend reversal (Stadt Zürich Statistik, 2023b). The advantages of living in the central city have been increasingly appreciated, and people want to settle here once again. This trend towards urbanization has created an insatiable demand and calls for densification. For this reason, since the 2010s, high-rise buildings have been increasingly constructed, allowing more people to be accommodated in a small space.

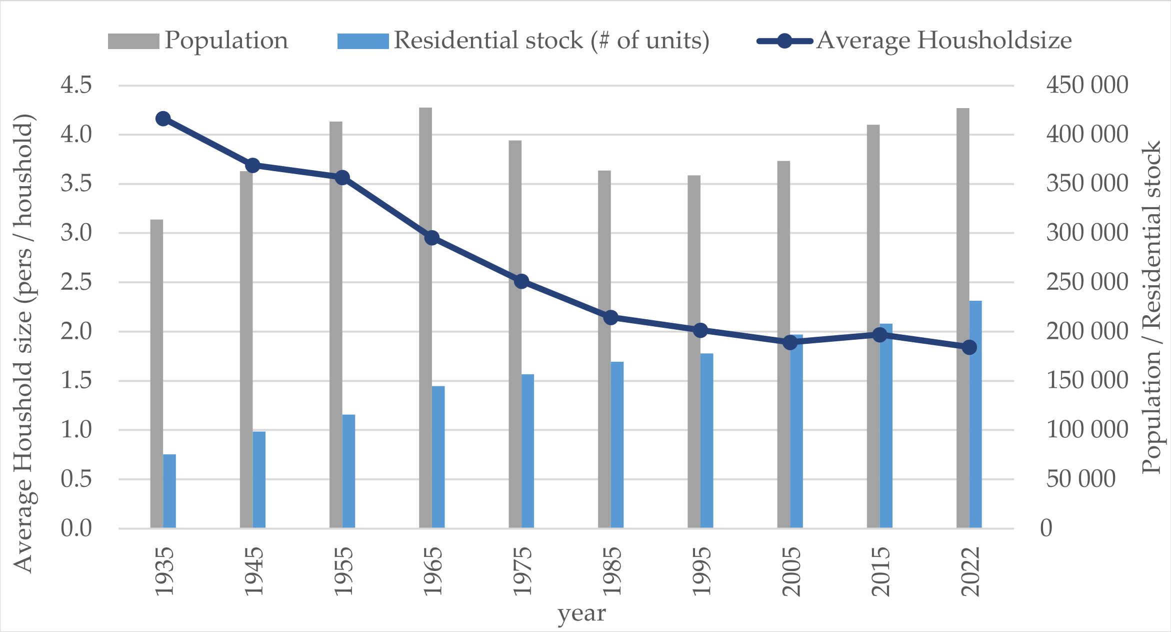
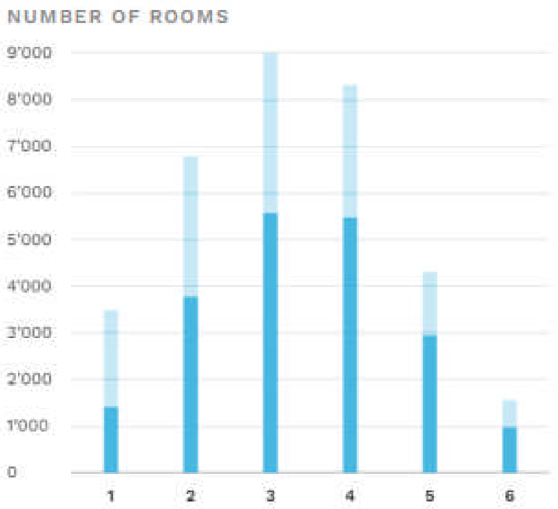


Figure 2.2 Stock of residential units, population and resulting average household size in Zurich-City, 1935-2022. Chart by Chantal Meier, based on *Federal Statistical Office et al., 2022; Stadt Zürich Statistik, 2023; Stadt Zürich Statistik & GWZ, 2023*

It is noteworthy that the almost equally high population in the year 1960 could be accommodated in significantly fewer housing units. In 1965, there were 144,659 units, whereas as of 2022, there are 231,522 units (Stadt Zürich Statistik & GWZ, 2023), that refers to an increase of 60%. This can be explained by the change in average household size, which has steadily decreased in recent years, resulting in fewer people living in one apartment: While in Switzerland, in 1930, more than half of households had more than 5 occupants, today (2021), almost half of households are either 1- or 2-person households (Federal Statistical Office et al., 2022). This trend is even more pronounced in cities, which are increasingly inhabited by singles and couples. In more rural areas, more families are settled, which in turn slightly increases the average household size in less urban areas. This is clearly visible, when comparing the average household sizes throughout the canton of Zurich. The city of Zurich has by far the smallest household size with 1.98 persons per household (median: 2.31 persons) (Kanton Zürich, 2023). The fact that the high demand cannot be met is consequently also due to a significantly larger space requirement, as smaller households live in the same apartments, bigger households used to live in. In this context, market experts refer also to a "prosperity crisis" rather than a housing crisis (Ménard, 2023).

However, even with the increased space requirement, it remains undisputed that demand far exceeds supply. The company Realmatch360 compiles analyses based on search subscriptions on popular housing platforms like Homegate or Immoscout. These provide a good indication of how many people are currently searching for an apartment in a municipality. Realmatch360 distinguishes between Specific and Perimeter Seekers. Specific Seekers have a search subscription for Zurich-City. Perimeter Seekers search within an expanded perimeter. This is possible because when creating a search subscription, a desired municipality and an extended radius of X kilometers can be entered for searching as well. Therefore, if a person is searching in a nearby municipality plus a radius of X, including the municipality of Zurich, they belong to the Perimeter Seekers (Realmatch360, 2023). As illustrated in Figure 2.3, there are 10,654 Specific Seekers alone for Zurich-City. When adding Radius Seekers to the equation, the demand increases to 19,109 individuals. Since most seekers are interested in 3- or 4-bedroom apartments, despite the majority of households being 1- or 2-person households, the increased space requirement can also be observed in the demand. As this demand meets a supply of 161 units (Federal Statistical Office, 2022), of which 135 are for rent and 26 for sale (Realmatch360, 2023), it is clear how drastic the demand exceeds the supply.



NUMBER OF ROOMS	SPECIFIC		PERIMETER	
1	13.2 %	1'411	24.6 %	2'076
2	35.5 %	3'778	35.6 %	3'013
3	52.3 %	5'571	40.3 %	3'406
4	51.4 %	5'480	33.6 %	2'842
5	27.7 %	2'946	16.1 %	1'359
6	9.1 %	969	7 %	588
Total		10'654		8'455

Figure 2.3 Distribution of demand based on searching descriptions on real estate platforms in Zurich-City (Realmatch360, 2023)

2.4.4 Resulting Dynamics

This significant demand surplus leads to an extremely competitive market with various consequences. So have rents throughout the city increased substantially in recent years. According to statistics from the Zurich City (Stadt Zürich Statistik, 2023e), rents have increased by almost one fourth (24.3%) since 1993 (see Figure 2.4). With another methodology the development can even be referred to as an increase of 40% during the last 20 years (Frey & Stadt Zürich, 2022). Besides the high demand, often the high booming economy is named as a further cause for rent increase as it attracts high earning professionals to the city with a great willingness to pay. This promotes an increasing supply of new and high-priced housing at the expense of old low-cost housing that is demolished for this purpose (Vakaridis, 2023).

Increased rents and the increasing dwindling of affordable housing then in turn leads inevitably to gentrification and the displacement of certain population groups. To counteract this gentrification and maintain a diverse city population, the city has set a goal to make one-third of all apartments nonprofit by the year 2050, offering them at cost-covering rents instead of market rents. This is to be achieved through various means, such as the Housing Fund created in 2023 with CHF 300 million. The funds from this program are intended to financially support buyers of multi-family properties when they rent out the apartments at cost-covering rates (Metzler, 2023).

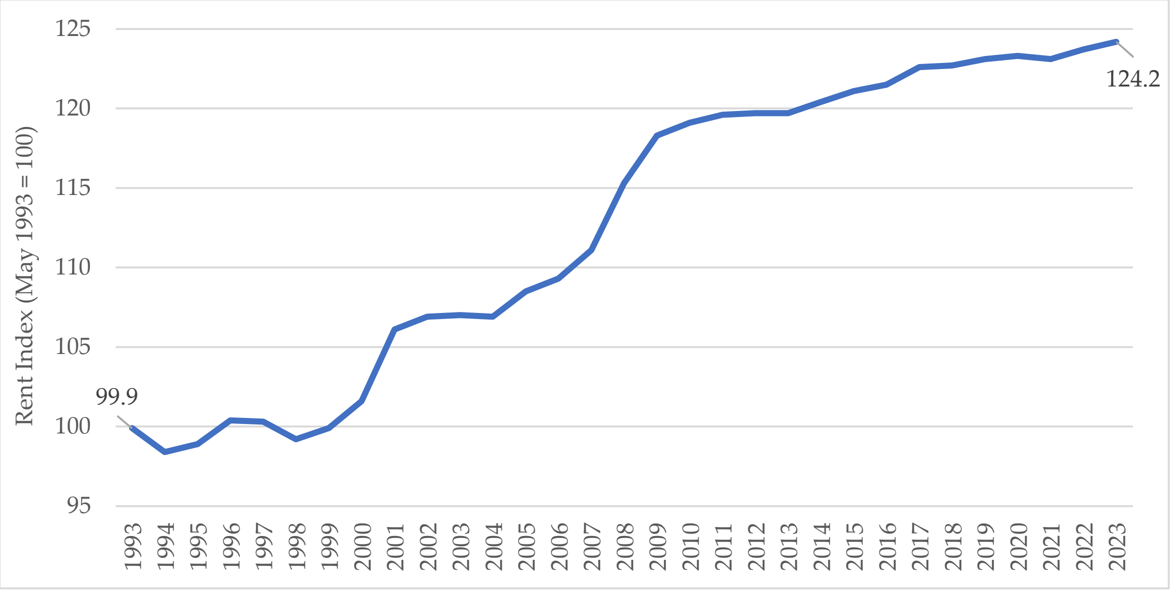


Figure 2.4 Indexed development of Rents in Zurich-City, 1993 – 2023 Chart by Chantal Meier, based on (Stadt Zürich Statistik, 2023e)

The scarcity also affects the dynamics between landlord / facility management and potential tenants. When an apartment is advertised, property management companies are often confronted with a flood of applicants (Aeberli, 2023) or long lines for apartment viewings. As various newspapers have reported on, especially an offer of a low-cost apartment, generally leads to a queue of hundreds of meters, which means that most interested parties do not even get an application form (e.g. Tagesanzeiger, 2019). To keep the influx as small as possible, some property management companies often employ the following trick: they keep the listing online for only a brief hour and then offer a viewing only to the applicants gathered in the short time span at a given time. (Aeberli, 2023). This makes the search especially hard for people that have not a fully flexible schedule to check permanently adverts and be free to attend one given appointment (Aeberli, 2023). In this context, the search for an apartment in Zurich-City is sometimes referred to as a "full-time-job" (Aeberli, 2023; Strobel, 2013). In addition, the application for the flat is becoming more and more extensive: The interested parties have to disclose a large amount of personal information (Aeberli, 2023) that goes beyond the "standard" excerpt from the debt collection register and even advertise themselves as if they were applying for a job, for example by including a letter of motivation as part of the new standard (Strobel, 2013). The great competition can also lead flat seekers to desperate means, such as voluntarily (and illegally) offering to pay an administration even more than the advertised rent (Meier, 2023) or offering other gifts (Strobel, 2013).

Interestingly, despite the low vacancy rate, a comparatively high rate of apartment turnover can be observed. As in Switzerland, every move is associated with a mandatory registration process, statistics are available for relocations. In this context, Rey (2020) found that despite the low vacancy rate, there's a high rate of apartment turnover in the city: On average, about 2,300 people move each month within Zurich City (Rey, 2020). Since Switzerland has certain regular termination dates when an apartment can be terminated with a 3-month notice without the need to find a subsequent tenant, this number is higher even higher during April and October, the months colloquially termed "moving months," amounting to about 2,800 individuals (Rey, 2020).

3 Data

The following section aims to give a brief overview on the extend and origin of the used datasets. The thesis is mainly based on two datasets combining real estate listings from different and commonly used online platforms.

These two datasets originate from Meta-Sys AG[3] (hereafter: Meta-Sys), a company formed in 2002 and specialized in data crawling. The data did not come directly in raw form after the crawling process but have undergone the typical processing conducted by Meta-Sys. This processing is described in the following section 3.1, as understanding it is crucial for both the methodology and interpretation of the results. The source is not always explicitly mentioned in this paragraph, but unless otherwise indicated, it is consistently the company of Meta-Sys and corresponds to either provided documents on the crawling process or information obtained through various discussions.

In addition to the real estate data obtained through Meta-Sys, various statistics related to the real estate market or real estate-related topics are also important (see section 3.2). These statistics originate either from the City of Zurich (e.g. Stadt Zürich Statistik, 2023a, 2023d, 2023b) or the Federal Statistical Office (FSO) (e.g. Federal Statistical Office, 2022; Federal Statistical Office et al., 2022) and have partly already been incorporated in the background section (see 2.4) to provide a more comprehensive overview of the Zurich real estate market.

3.1 Crawled Real estate listings

The thesis is mainly based on two datasets with residential real estate advertisements in Zurich-City and its neighboring communities. One dataset contains historical entries that were published online between 2004 – 2017 and the second dataset contains more recent data from 2017 – 2022. Except for the different time periods the two datasets do not differ at all as they are equally structured and cover the same variables. This makes it simple to combine them in a statistical tool. The last entries from the recent dataset are dated on December 23rd, 2022, that can be considered the reporting day, while the first entry in the historical dataset is from January 1st, 2004, that can be considered as the starting day. They do not only cover Zurich-City but also all its official neighboring communities, that are namely the following (listed in alphabetical

[3] https://www.meta-sys.ch

order): Adliswil, Dübendorf, Fällanden, Kilchberg, Maur, Oberengstringen, Opfikon, Regensdorf, Rümlang, Schlieren, Stallikon, Uitikon, Urdorf, Wallisellen, and Zollikon.

The two datasets were received freely for study purposes from Meta-Sys. They have collected all online listings for residential real estate (to rent and buy) since 2004 from the various online available platforms. The two sets contain a total of 421,602 listings for the Zurich City and the neighboring communities. Most of the advert listings are described in German, a few English descriptions can be detected too.

For this thesis not all collected variables (see Table 3.b) are relevant. The most important ones are the texts title and description, as they build the source of unstructured text to perform the named entity recognition on, and the corresponding geocoordinates for toponym resolution. Below follows a typical title and description, chosen randomly from the dataset, in order to provide an example (translated from German into English, relevant manually detected location references marked in **bold formatting**):

Title	4 ½-room garden flat
Description	In the **Frohbühlpark** in **Zurich-Seebach**, directly on the **northern city border**, we are selling by appointment a spacious, comfortable 4 1/2-room flat on the ground floor, with a large, sunny seating area and additional backyard. The flat has a modern finish, with a beautiful kitchen with lots of storage space, high level oven, glass ceramic, dishwasher. Bathroom with double washbasin and large mirrored cabinet, separate guest toilet with shower. All rooms have parquet floors. A washing machine and a tumble dryer are located in a separate room, directly in the flat. The flat is currently rented. Are you interested? Ask for our documentation or arrange a viewing appointment with us.

Table 3.a Example of text and description from listing to buy, located at Schaffhauserstr. 641, 8052 Zurich

3.1.1 Preprocessing steps of data

The whole dataset goes through various processing steps, conducted by Meta-Sys. Their crawling process is schematically described in Figure 3.1. The process includes the crawled data volume ('spider volume'), which is stored in a database (*Spider Database (Operations))*. The *Adscan Database (Operations)* constantly checks with the spider volume whether an advertisement still exists online and whether new ones have been added, which are then added to the database. Afterwards, the data goes through various processing steps until it reaches the new *AdScan Database (Warehouse).* From this database, exports are made to customers who purchase data from Meta-Sys or access it directly via an API interface. The datasets used in this thesis also originate as an export on December 23rd, 2022 (reporting day) from this database.

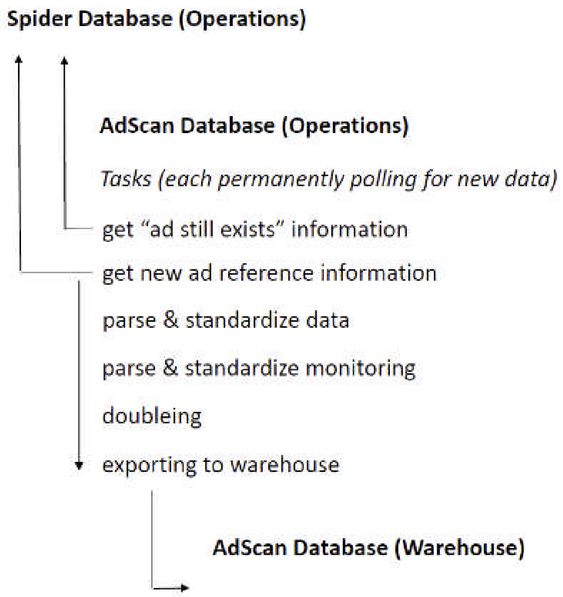

Spider Database (Operations)
AdScan Database (Operations)
Tasks (each permanently polling for new data)
get "ad still exists" information
get new ad reference information
parse & standardize data
parse & standardize monitoring
doubleing
exporting to warehouse
AdScan Database (Warehouse)

Figure 3.1 Crawling and Processing steps of Meta-Sys. Source: Metasys, 2023

Between the two databases *AdScan Database (Operations)* and *AdScan Database (Warehouse)* several steps of processing are taken:

1) *Parse and Standardize*

The parser extracts raw information from an HTML file (namely, various real estate platforms). Standardization generates uniform information for each field (see also Table 3.b). At this stage, the content of the information is not checked but is simply represented correctly. An exception is the objecttype, which is already plausibilized at this stage via text description.

Further, this step includes the linkage of geocoordinates to the addresses:

Addresses are compared with the Meta-Sys address database. If a match is found, the geocoordinate is taken from there. If no match is achieved, the alternative is to use the

Google Geocoding API. If the advertisement does not contain a street name, the georeference is created at the postal code level. This difference between precisely assignable coordinates and approximately assigned coordinates is also recorded (see Table 3.b).

2) *Parse and Standardize Monitoring*

When a platform changes its structure or a new platform is added to the crawling process, a test is conducted until all the following questions can be answered with 'Yes':

- ✓ Has all relevant information been parsed?
- ✓ If relevant information is missing, is it missing on the website itself and not due to the processing process?
- ✓ Is the standardization of the parsed information being done correctly?

3) *Duplicate Cleansing*

In the process of duplicate cleansing, the primary objective is to address situations where the same advertisement may be simultaneously listed on multiple platforms to reach a broader audience. Additionally, there are cases where advertisements briefly disappear from the platform and are subsequently re-added. To prevent the second addition from being treated as a new advertisement, as this influences also further variables, Meta-Sys consolidates such instances. Such an influenced variable would be the marketing period, that refers to the numbers of days an advert is online and is in the real estate branch commonly used as an important indicator for demand. For the downstream users of the data, this consolidation is typically a disadvantage, which is why Meta-Sys cleanses these duplicates. This process involves several steps and is carried out within the same postal code area. To perform the comparison, objects with the same street, street number, area, and floor are initially compared. The floor is deemed significant because the "same" apartment in the same building on different floors is not considered a duplicate advertisement according to Meta-Sys. In the second step, the duplicate filter calculates the co-similarity between the two advertisement texts. If this similarity exceeds a certain value, the two pairs are recognized as duplicates and merged into a single entry.

The crawling process and the different steps of data post-processing and added information through Meta-Sys result in the variables as displayed in the following table 3.a.

Name of Variable	**Meaning**	**Example of *Value* / Code and meaning**	**Coverage[4] (%)**
id	Number of identification	*40340098*	100%
double1group_id	Number of ID for grouped objects (same advert, different platform)	*34959974*	100%
mpname	platform advert was published	*Scout24*	100%
te305o_tenure_id	Category of offer	413 (offer to rent) or 414 (offer to buy)	100%
found_date	First found date	*26.08.2017*	100%
last_found_date	Last found date	*28.08.2017*	100%
s_title	Title of listing	*Attraktive Wohnung an zentrumsnaher Lage*	97.7%
s_description	Description of listing	*Im Auftrag unseres Kunden verkaufen wir in der Überbauung "Am Pfingstweidpark" im aufstrebenden City West, eine schöne, moderne 3-½-Zimmerwohnung. Verkaufsrichtpreis: CHF 1'100'000. Haben wir Ihr Interesse geweckt? Gerne senden wir Ihnen die ausführliche Verkaufsdokumentation über dieses interessante Angebot zu.*	98.6%
s_construction_year	Construction year of building	*1937*	45.3%
s_zip	Zip Code	*8003*	99.6%
s_city	Municipality	*Zürich*	99.9%
s_street	Streetname	*Badenerstrasse*	93.7
s_stnr	Street number	*343*	82.1%
te302_objecttype_main_id	Main Category of object	430 (house) 445 (apartment)	100.0%
te302_objecttype_id	More specific category of main object	430-439 (different categories of houses) 445-457, 528 (different categories of apartments)	100.0%
s_surface_usable	Usable floor space (sq m)	*96*	71.1%
s_surface_living	Living space (sq m)	96	85.3%
s_surface_property	Surface area of property (sq m)	579	2.5%
s_rentextra	Service charges / Month	369	67.7%
s_grossrent	Gross Rent / Month	*402*	72.6%
S_netrent	Net Rent / Month	*3920*	72.0%
S_salesprice	Price of Sale	*850,000*	6.4%
Te304o_currency_id_(…) *(…) = rentextra / grossrent / netrent / salesprice*	Currency of corresponding value (rent, price)	415 (CHF), 416 (by request), 417 (USD), 418 (EUR)	86.7% / 80.5% / 88.3% / 7.1%
Te306o_typeofprice_id_(…) *(…) = rentextra / grossrent / netrent / salesprice*		407 (other), 408 (annual), 409 (annual / s qm), 410 (per month), 411 (by request), 414 (selling price / sq m)	86.2% / 80.4% / 87.5% / 7.1%
S_nbrooms	Number of rooms	*4.5*	95.4%
S_floor	Number of floor	2	80.9%
Te307o_floor_id	Special type of floor	200 (lower ground floor), 404 (attic floor), 406 (raised ground floor)	<0.1%

[4] Share of all listings (421,602) that show this information

Name of Variable	**Meaning**	**Example of *Value* / Code and meaning**	**Coverage[4] (%)**
te_a_sicht_oc te_a_ofen_oc te_a_balkon_oc te_a_wintergarten_oc te_a_garten_oc te_a_lift_oc te_a_rollst_oc te_a_wasch_oc te_a_neu_stand_oc te_a_minergie_oc te_a_autoab1_oc te_a_autoab2_oc	Different variables on the amenities the apartment does provide (view, oven, balcony, winter garden, garden, elevator, wheelchair accessibility, washing machine, first-time occupancy, mingergy, parking space included in rent, parking space not included in rent)	All these amenities show the categories: (-1) = unknown / no information 0 = not existing 1 = existing Some have further numbers (2-5) indicating more specific information on the type of amenity within the category. As for example balcony (e.g. terrace, large balcony, small balcony, etc.) or view (e.g. lake view, mountain view, etc.).	24.1% 10.4% 66.3% 1.6% 13.0% 33.9% 5.3% 5.7% 12.7% 3.0% 24.3% 37.8% *(value -1 unknown excluded)*
Double1group_startday	Start of double group	*26.08.2017*	100.0%
Double1group_endday	End of double group	*28.08.2017*	100.0%
Te308o_geoprecision_id	Precision of geocoding	177, 185 (exact address, different processing), 187 (address interpolated), 179 (street), 180 (specific object), 181 (zip code), 182 (community)	98.7%
S_lon	Longitude (WGS 84)	*8.525647*	99.0%
S_lat	Latitude (WGS 84)	*47.41653*	99.0%
url	Original Link of Listing	*https://www.homegate.ch/kaufen/106796622*	95.6%
Metasysid	Internal hash meta-sys	*0d0848eb707145d095ff3dfd16d1c12eefc645d9*	50.9%
te303o_availability_id	Category of availability	89 (per immediately) 90 (per arrangement), 92 (fixed date)	95.5%
s_availability	Date of availability	*01.09.2017*	65.9%
bfscode	FSO Code of municipality	*261*	100%
checked_today	Availability checked on this day	0 = no, 1 = yes	100.0%
Insperiod	Time period listing was online (days)	*79*	100.0%
Postdouble	Double Cleaning by Meta-Sys among the different platforms	0, 1	100%

Table 3.b Overview on available variables in the dataset and their coverage among all the entries, coverage calculated by Chantal Meier

Fortunately, the text and description of a listing can be considered as one of the core elements of a listing and is therefore widely available in the given dataset as 97.7% of all the listings have a title and 98.6% have a text description. In contrast, data related to additional attributes such as the construction year, surface or available amenities such as balconies and views are considerably scarcer. Complete coverage, at 100%, is achieved only for variables introduced by Meta-Sys themselves or those that are absolutely mandatory for the user to enter when creating a listing, such as specifying whether it is for rent or purchase.

3.2 Data by City of Zurich / Federal Statistical Office (FSO)

A second important data source for this thesis are official statistics published by either the City of Zurich or the Federal Statistical Office (FSO). Many of them have already been used to describe the residential real estate market in Zurich-City in the Background section (see 2.4).

Often, these statistics originate from the same survey but are made available for public download at different locations and with varying levels of detail. For example, statistics on vacancy rates can be obtained from the Federal Statistical Office for the entire Switzerland at the municipality level, while the City of Zurich provides the same statistics in greater detail for the various city districts within the municipality.

The used material from the administrative entities can be divided into 3 categories:

1) **Data on existing stock / data on offer**: Equally important are the statistics on the number of vacant dwellings and the vacancy rate. On one hand, these serve as a crucial indicator for the general description of the real estate market, as discussed in section 2.4. On the other hand, they can also be used for comparison with Meta-Sys data, for example, to assess how many of the available apartments in the market can be found in the Meta-Sys dataset.

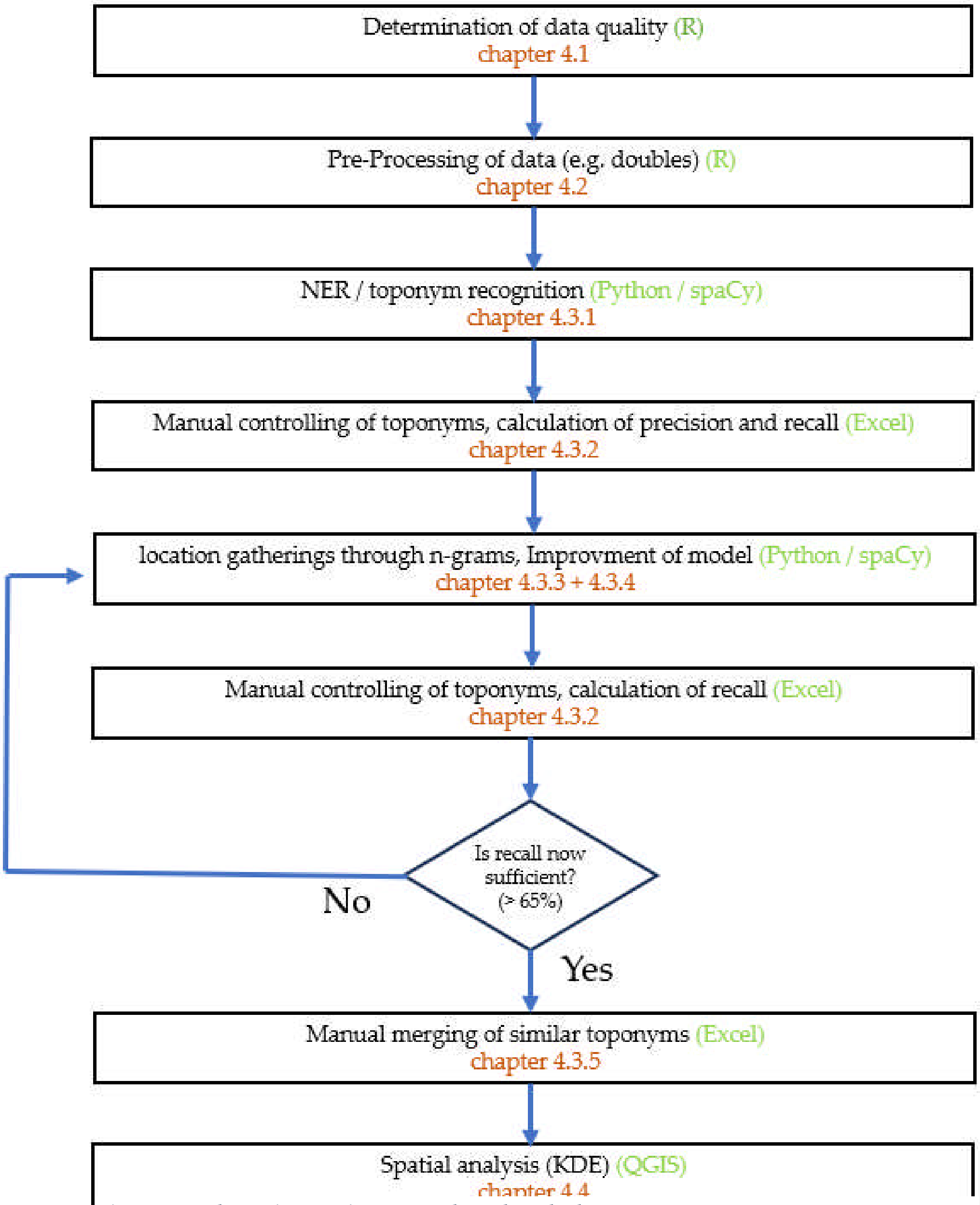
2) **Demographic data / data on demand**: The utilized statistics include demographic data, such as population figures. These data illustrate the periods of urban flight in the 1970s/80s up to the current trend reversal, where increased population growth has occurred since the urbanization around the turn of the millennium. Contrasted with statistics on the number of residential units in the city (also referred to as residential stock), it is possible to make statements about household size and its changes.

3) **Geographical data**: Lastly, the precise official city division is also incorporated into the utilized data. In order to process the extracted toponyms, especially the neighborhood designations, to toponym resolution and compare them with the officially existing city quarters, the official gazetteer of the City of Zurich is consulted. This gazetteer includes the boundaries of districts and city quarters, as well as the smaller statistical quarters, which are named after significant locations. All geographical levels are available as shapefiles, simplifying the comparison with the self-generated spatial footprints.

4 Methodology

This chapter is intended to provide a chronological overview of the processing steps that the data have undergone and the methods that have been applied. In the graphic above (see Figure 4.1**Fehler! Verweisquelle konnte nicht gefunden werden.**), the process flow and the programs used are schematically displayed.

For the analysis the data has been processed and analyzed partly in R (preparation and description of data), python (natural language processing with spacy and analysis of n-grams), Microsoft Excel (controlling and improvement of toponym list) and QGIS (spatial analysis of recognized toponyms through kernel density estimation).



Determination of data quality (R)
chapter 4.1
Pre-Processing of data (e.g. doubles) (R)
chapter 4.2
NER / toponym recognition (Python / spaCy)
chapter 4.3.1
Manual controlling of toponyms, calculation of precision and recall (Excel)
chapter 4.3.2
location gatherings through n-grams, Improvment of model (Python / spaCy)
chapter 4.3.3 + 4.3.4
Manual controlling of toponyms, calculation of recall (Excel)
chapter 4.3.2
Is recall now sufficient? (> 65%)
No
Yes
Manual merging of similar toponyms (Excel)
chapter 4.3.5
Spatial analysis (KDE) (QGIS)
chapter 4.4

Figure 4.1 schematic overview on conducted methods

4.1 Description of data

As seen earlier, the Zurich real estate market grapples with an overwhelming demand, which coupled with a very limited supply (see section 2.4). This scenario undeniably impacts the dataset under scrutiny, given its intimate linkage to the real estate market dynamics. Therefore, it is pertinent to assess the extent of this influence.

Initially, the dataset is intrinsically limited. In another municipality of a comparable size, yet with a greater number of apartment listings, the dataset would invariably be more expansive. This scenario prompts inquiries about the potential influences on the extant data. It is worth noting that the dataset might be even more limited than it appears, as not every listing finds its way online. Industry insiders often posit that around 70% of Zurich apartments are never advertised online, instead being passed on through "informal" channels. Moreover, the intense demand, coupled with the substantial administrative tasks for property management, might shape the content and nuances of the property descriptions..

Concretely, these assumptions are captured in the following two questions:

a) *How many apartments are rented off-market and do never appear in the database?*

b) *Is there a tendency towards "poor descriptions" in a highly sought-after market?*

4.1.1 Comparison with official real estate data (stock, vacancy, movements)

The first question is addressed with the following approach:

a) *Comparison with official stock, vacancy and movements*

In light of the prevailing conditions, it is plausible to infer that Zurich's pronounced housing demand precipitates a sizable chunk of "off-market" apartment transactions. This, consequently, equates to fewer online listings and, by extension, a diminished dataset − an observation that stands even without in-depth exploration. Nevertheless, pinpointing this discrepancy remains paramount. Drawing from Chen and Biljecki's (2022) methodology, where they juxtaposed crawled listings against the extant housing stock to determine the listings' representation, we have the opportunity not just to reference the housing inventory but also to directly contrast with the FSO's annual vacancy statistics, released every June 1st. These statistics elucidate both the total vacant units and the relative vacancy rate.

To discern the listings available online each year from 2004-2022 on June 1st, filters were applied to the variables Double1group_startday and Double1group_endday.

As delineated in section 2.4.4, mandatory registration accompanies each relocation, making relocation statistics accessible. Rey’s (2020) research underscores the peculiarity of Zurich's housing dynamics: despite a dwindling vacancy, the city observes a substantial apartment churn. A monthly average of 2,300 inhabitants resettles within Zurich City, peaking to around 2,800 during the April and October "moving months" (Rey, 2020).

Furthermore, the city annually releases data on intra-city relocations (Stadt Zürich Statistik, 2023d) as well as migrations to and from the city (Stadt Zürich Statistik, 2023f). An aggregation of intra-city relocations and inbound migrations should afford a reliable estimate of annual new apartment occupants in the city. It's postulated that apartments vacated by outgoing residents are either inhabited by newcomers or those moving within the city. The marginal number of apartments that go unoccupied and thus contribute to the vacancy rate are consequently overlooked.

However, since both statistics refer to the number of people, no conclusion can be drawn yet on how many apartments are occupied by these individuals. An approximation for this could be derived by accounting for household size (Federal Statistical Office et al., 2022): if the number of relocating or incoming individuals is divided by the average household size for the given year, this then results in the number of households relocating or moving in, which can be equated to the number of apartments newly occupied in the city.

4.1.2 Correlation between description lengths and vacancy rate

The second question is addressed with the following approach:

b) *Correlation between descriptions lengths and vacancy rate*

To investigate whether the competitive Zurich housing market could lead to poor text descriptions, first, a method must be established to identify a poor description. One could simply assume that poor text descriptions are shorter and consequently consist of fewer words compared to substantial descriptions.

Since the primary hypothesis is that in Zurich-City, listing descriptions may be shorter, the average number of words per description is examined for each municipality. Given that the dataset includes not only Zurich municipalities but also neighboring communities, the latter can be included for analysis. Then vacancy statistics can again be referenced. In a subsequent step, it can then be assessed whether there is a correlation between the average number of words per listing description and the average vacancy rate according to FSO.

The thereby calculated correlation value falls within the range of -1 to +1, where 0 indicates no correlation, 1 represents a complete or perfect correlation, and -1 signifies an absolute lack of correlation (Akoglu, 2018). In more detail, the ranges from 0.1-0.3 refer to a weak correlation, 0.4-0.6 refer to a moderate correlation, and 0.7-0.9 refer to a strong correlation. The same ranges can be applied in the negative range (Akoglu, 2018). The calculated value is categorized within the previously described range, which allows to answer the question of whether a lower vacancy rate (as an indicator for a highly sought-after market) leads to shorter (as an indicator for poorer) descriptions and thus fewer potential toponyms to detect.

4.2 Preprocessing of data for location recognition

First, the two datasets (historical and recent data) must be combined to make an overall-analysis which is possible through a simple "bind-rows" function in R as they show the exact same structure. Afterwards, the pre-processing of the data can be started.

4.2.1 Rough pre-processing and filtering of relevant variables

Listings with no title and no text description can be excluded directly, as they cannot be mined for toponyms and therefore have no utility. Additionally, the community of Zurich is extracted using the unique FSO code assigned to every municipality in Switzerland. It is advisable to use the FSO code instead of the municipality name, as variations in spelling and potential errors, especially with the letter "ü" in the German spelling of Zurich ("Zürich" or alternatively "Zuerich"), can lead to discrepancies.

Fortunately, the FSO code has already been correctly assigned to all municipalities in the dataset, despite variations like misspellings, the complete absence of community names, different languages (particularly German and English), or alternative spellings and terms (Zürich-City, Zurich, Zuerich, etc.). To optimize processing speed on the relatively limited-capacity computer, all columns that are not of immediate interest are excluded upfront. If needed later, these columns can easily be reintroduced using the unique ID assigned by Meta-Sys to each listing.

Although Meta-Sys has added geocoordinates to all listings, the precision of these coordinates varies. This variation arises because not every apartment or house is published with complete or accurate address details, resulting in geocoordinates that do not pinpoint exact locations. For instance, if a listing includes only minimal information required for user entry, it may contain only a zip code and community, with no street name or number provided. Since the gathered information will be later analyzed within a spatial context, it is logical to exclude listings with "imprecise" geocoding that can be readily identified. Such information is indicated in the 'Te308o_geoprecision_id' column (see Table 3.b), where codes 177 and 185 correspond to exact address information and, consequently, precise geocoding.

4.2.2 Identification of duplicates

As described in section 3.1 the data has already undergone a cleaning process by Meta-Sys where the same listing on different platforms has been merged into one duplicate group through the variable *double1group_id*. This does, however, not avoid all duplicates that should be considered for the given analysis.

As indicated by Chen and Bilijeck (2022) it is a common strategy to delete and re-upload a listing as the ones added more recently will be displayed more popularly on the platform and reach consequently a broader audience. For this case (re-upload) the duplicate filter of Meta-Sys can already detect the double if the re-upload has been made within a short time period. If, however, the re-upload has a delay the double is not identified and counted as a separate listing.

Moreover, it is plausible that the same listing has been published several times over the investigated time period because the apartment / house could be rent or sold multiple times. This seems likely, as according to Rey (2020) the average stay in a private apartment is 7 years while for non-profit apartments it refers to 14 years. So, especially rental units can be expected to show up more than once in the dataset. This is for the given analysis not necessarily a problem if the location description of the listing changes as well (see also Chen & Biljecki, 2022). If, however, the landlord or property management remains the same for the several rent-outs they might use the same description multiple times.

The same is the case, when multiple apartments within a building are listed, using the same description. These are not counted as duplicates because the Meta-Sys duplicate filter compares not only text and address but also the floors to avoid losing a listing in such a case.

For the given analysis it does not make sense to gather the same location references within the same address twice as this would mean to count the retrieved location of one listing creator twice. In this context, it is irrelevant whether the same description repeats after a re-upload due to a rental strategy or a routine tenant change, or whether it pertains to a description for multiple units at the same location. Consequently, the data-preprocessing must include further identification of doubles and shall exclude all adverts that refer to the same address and show an equal title and description by comparing text information for the same address.

4.3 Named entity recognition / location recognition through spaCy

Currently, there is different software on the market that is suited to perform NER tasks and is primarily open-source and free to use. The 5 most popular ones are namely StanfordNLP, NLTK, OpenNLP, SpaCy, and Gate (Schmitt et al., 2019). Overall, it is difficult to determine which one of them performs best because they have led to different results in different studies and are barely for the same task and methodology directly compared to each other (ibd.).

Schmitt et al. (2019) addressed this lack of comparability with their study and tested five NER software tools along with two corpora. They concluded that StanfordNLP performed the best, although this was primarily for the entity type "Organization." For "Persons" and "Locations," there was hardly any difference, making it challenging to select the best NER program for this thesis, especially since a comparable study does not seem to exist for the German language.

Therefore, for this thesis, I made the decision to conduct the recognition of used placenames in real estate listings with spaCy, that can be used with R or Python, which I did at that point already have most experience with.

4.3.1 Processing steps in spaCy

spaCy is an open-source Python library that offers advanced capabilities for Natural Language Processing (NLP) at no cost (spaCy, 2023). Spacy offers the following features, that are partly available on their own, but often used in combination during an NLP task (spaCy, 2023, see also Figure 4.2):

Feature	**Description**
Tokenization	Segmenting text into words, punctuation, marks, etc.
Part-of-speech (POS) Tagging	Assigning word types to tokens like verb or noun

Lemmatization	Assigning the base forms of words
Sentence Boundary Detection (SBD)	Finding and segmenting individual sentences
Named Entity Recognition (NER)	Labelling named "real-world" objects, like persons, companies or locations
Entity Linking (EL)	Disambiguating textual entities to unique identifiers in a knowledge base
Similarity	Comparing words, text spans and documents and how similar they are to each other
Text Classification	Assigning categories or labels to a whole document, or parts of a document
Rule-based matching	Finding sequences of tokens based on their texts and linguistic annotations, similar to regular expressions
Training	Updating and improving a statistical model's predictions
Serialization	Saving objects to files or byte strings

Table 4.a Overview on spaCys capabilities and their meaning, quoted from (spaCy, 2023)

For an NLP task, spaCy offers trained pipelines that can be installed as python packages (see Figure 4.2). They contain all the necessary steps to conduct a NER, including NER, and could also be enlarged with further steps as e.g. implementation of own training data, when required by the user (spaCy, 2023).

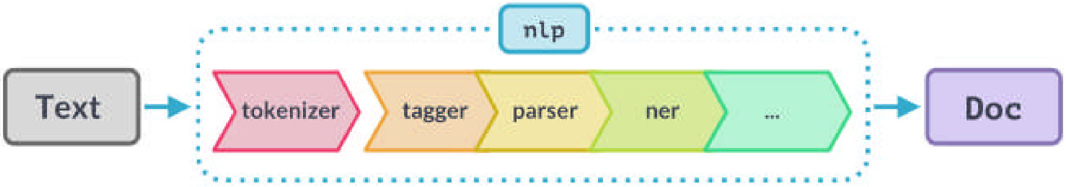


Figure 4.2 Pipeline for an NLP in spaCy (spaCy, 2023)

First, a *tokenizer* breaks the analyzed text down into tokens. This is basically for every NLP task the first necessary step that needs to be taken as it means to identify and segregate the basic units that will be analyzed in the further steps (Webster & Kit, 1992). This may sound straightforward, but when considering that a token is by no means always just a single word, which would be easy to determine, but can also be a sequence of words (e.g., in an expression), it quickly becomes apparent why the task is not as simple as it may seem (Webster & Kit, 1992).

Then, tagging, short for part-of-speech tagging, is processed. During part-of-speech tagging, also referred to as grammar-tagging, sentences or words are put into a context to the other sentences and words within the same paragraph (Chiche & Yitagesu, 2022). This usually includes a grammatical investigation and assigning the words to the different syntax groups such as nouns, verbs, or adjectives, etc. which shall reduce semantic ambiguities in the further processing steps of an NLP (Chiche & Yitagesu, 2022). This would reduce the number of candidates for toponyms drastically, as place names could only be nouns.

The next and final step before the actual NER can be performed is the parser. This step supplements the identified syntax with their dependencies, thus determining the relationships between individual words (spaCy, 2023).

The last, but most crucial part for the given task, is the execution of the NER itself. It is capable of detecting the following entities for the German language (spaCy, 2023):

- Persons (tagged as PER)
- Locations (tagged as LOC)
- Organizations (tagged as ORG)
- Miscellaneous (tagged as MISC)

Since the goal of this thesis is the extraction of place names, only the entity typ LOC will be kept and further analyzed. After the previously described text processing steps have been executed in Python, the identified toponyms can be evaluated to determine what improvements still need to be made. This can be made through precision and recall.

4.3.2 Precision and Recall (first round)

Typically, the performance of a NER system is evaluated using the metrics of precision and recall. These metrics were introduced during the Information Extraction (IE) tasks of the Message Understanding Conferences (MUC) and have proven to be effective evaluation standards that continue to be used today (Grishman & Sundheim, 1996).

Adapting Grisham & Sundheim (1996) on the given case of toponym recognition, the formula for precision and recall would be the following (Ettlin, 2011):

$$precision = \frac{number\ of\ correctly\ tagged\ toponyms}{number\ of\ tagged\ toponyms}$$

$$recall = \frac{number\ of\ correctly\ tagged\ toponyms}{total\ number\ of\ toponyms}$$

To explain in words, for precision, the found toponyms are compared to each other. The more correctly tagged toponyms there are, the higher the precision. However, precision does not provide information about how many toponyms were found in general. To determine this, the recall formula is used. In this case, a sample of the investigated data must be checked in order to determine the total number of available toponyms, which are then compared to the correctly found toponyms. Again, the more correctly tagged toponyms, the higher the recall.

After an initial run with spaCy, the recognized toponyms are sorted in descending order based on their detected frequency and exported into an excel-file for manual post-processing. This allows for the determination of a weighted precision in the first step. This makes sense in the sense that toponyms that are rarely found are not suitable for a spatial analysis anyway, as the later-applied Kernel Density Estimation provides reliable results only when a sufficient number of georeferenced toponyms are available. For this initial sample, all toponyms with at least 30 occurrences are examined. This is because the number 30 is often cited as the minimum sample size for statistical purposes, even when the later processed Kernel Density Estimation was only made for entries with a significantly higher count.

For toponyms occurring more than 30 times, the list can then be manually reviewed, and for each toponym, it can be determined whether it represents a correctly labeled toponym or not. After this process the precision can be calculated among the controlled toponyms.

However, since only the frequently occurring toponyms will be of great importance, it makes sense to also calculate the weighted precision, as it is ultimately more informative. It is calculated as follows:

$$precision = \frac{number\ of\ correctly\ tagged\ toponyms * frequency}{number\ of\ total\ tagged\ toponyms}$$

In the second step, recall is determined using a random sample of listing texts. For this purpose, a dataset of 200 randomly selected listing texts is generated and exported into excel. From all the listing texts, manually detected toponyms are collected. These detected toponyms are then compared to the exported list with automatically detected toponyms via the formula vlook-up in Excel. This allows to calculate recall for the given sample.

4.3.3 Improvement of model

Since recall has been manually determined for all toponyms that will later be used for spatial analysis and answering the research questions, all incorrectly detected toponyms can be easily excluded. This is done by inputting them into a list as a .csv file and then excluding them for the recognized toponyms back in Python. This step allows perfecting recall to 100%, as no incorrect toponyms will be included in the analysis. This includes also toponyms that are correct but of no use for the analysis.

Simultaneously, work can also be done with the non-detected toponyms from the sample. These are also combined into a .csv list and reintroduced into Python. There, they serve as a small gazetteer, used as a lookup list. This way, these toponyms can also be detected in all other listings outside the sample set and included in the overall analysis. This automatically improves precision for the overall evaluation as well.

With the two provided lists (incorrectly detected toponyms to be removed and undetected toponyms to be added), the analysis can be conducted again. Subsequently, another sample set can be evaluated to determine the (hopefully) improved precision. Of course, care must be taken to ensure that none of the listings overlap with the previously selected sample, which is easy to verify since each listing has a unique ID.

If the precision is re-evaluated and is not sufficiently high, the newly detected toponyms can be added to the list that complements the NER results. This process is repeated until the precision can be deemed high enough.

4.3.4 Further location gathering through N-grams

The NER has shown that the term "Lage" (location) occurs very frequently. During manual inspection, it was also noticed that this often occurs in conjunction with a preceding adjective that describes this location (e.g., "an ruhiger Lage" - in a quiet location, "an sonniger Lage" - in a sunny location, "an guter Lage" - in a good location). Therefore, it was decided to include the various location descriptions in the analysis and extract them more effectively than through spaCy's NER.

For this purpose, so-called N-grams are suitable. These are often used in computational linguistics, for example, in tasks related to speech recognition, automatic language correction, or translation (Derungs & Purves, 2016). According to Purves and Derungs (2016 : 308) "N-grams are essentially look up tables, allowing estimations of probabilities of particular phrases (either in exact forms or represented by wild cards), generated from very large corpora". In their case, they build n-grams in the connection with the term near, as this often is used in combination with a toponym. So that the token that follows the term near is extracted in order to investigate spatial relations (Derungs & Purves, 2016).

For the term location a similar N-gram is built by extracting every token that is positioned in front of the term "location" (* *Lage*). The word *Lage* makes the formation of the n-gram simple, since it actually always occurs in the same form. With other German words, it may be that they are slightly different depending on the causus, which would then have to be taken into account. However, *Lage* in German actually always occurs in the singular and the same form, regardless of the case.

After the code has been completed with the detection of the n-grams, these can be added to the already discovered toponyms for further analysis.

4.3.5 Manual Merging of identical toponyms

In the toponym list, there are various similar spellings for one and the same toponym, but these are then evaluated as different entries. This is obvious for several reasons:

- Since the advertisement texts are user-generated content, it is possible that there are different spellings or spelling mistakes for one and the same toponym
- The German language also leads to the same word being detected several times, e.g. if it is placed in a different casus
- On the one hand, this may be due to the fact that spaCy segregates the place names in different tokens, for example, once a placename is recognized as a single word or once as part of a sequence
- The same toponym can be referred to with different place names, that obviously refer to the same location

These entries are manually reduced to the same description in Excel. The following Table 4.b shows an example of such a merge. One approach pursued is to consolidate different spellings or various grammatical cases into a single term. This is very easy to interpret manually. It becomes a bit more challenging with terms like "Zürichsee" (Lake Zurich). The term "See" (lake) is then attributed to Zürichsee based on the principle of what is most likely, even though, theoretically, it could also refer to another lake, such as Katzensee. According to Lieberman et al. (2010) this default-sense is one of two pre-dominantly used models and does assume it is always the most referred to location (e.g. Paris, France instead of Paris, Texas or in this case "Zürichsee" instead of the much smaller "Katzensee").

Detected Toponyms by spaCy	**Manually merged into**
Seefeld Seefeldquartier Zürcher Seefeld Zürich-Seefeld	Seefeld
Zürichsee Zürichsees Zürisee Gehdistanz zum See See	Zürichsee

Table 4.b Examples for manual toponym merging

4.4 Spatial analysis / toponym resolution

In the final part of the analysis, the collected toponyms containing at least 300 entries are spatially analyzed. This is relatively straightforward because each listing is provided with geocoordinates. Accordingly, a list can be extracted from Python, with each line containing the coordinates and the recognized toponym.

For vernacular geographies with no sharp boundaries density surfaces are particularly suitable as also used in the thesis of Hollenstein (2008) in Zurich-City. A classical choice for the creation of such density surfaces out of a point dataset is Kernel Density Estimation (KDE) (Grothe & Schaab, 2009). The kernel size is chosen very small scale (100m) as most location description refer to a small neighbourhoud and this way it is possible to create a detailed picture.

The resulting density maps for each toponym can then be categorized into classes based on their characteristics and spatial patterns, structuring the interpretation and description of the results. One class, for example, consists of the city neighborhoods and districts, as found in the official Gazetteer of the City of Zurich. Another group comprises entities that are frequently described due to their attractiveness or visibility, such as Lake Zurich or Uetliberg. For each class a few interesting examples are picked and displayed.

For the class of districts, additionally the official gazetteer of Zurich-City is used, to explore how the vernacular city districts differ from the official borders. For this purpose, the official city neighborhoods and districts, available as shapefiles, are also implemented in QGIS and compared with the footprints of the KDE.

5 Results and Interpretations

5.1 Description of data

5.1.1 Comparison with official stock / vacancy

Comparing the number of listings with the number of vacant units has proven to be of little relevance. It is readily apparent that the number of listings, filtered for the reference date on which the vacancy is measured, far exceeds it: On average there where 1,743 listings available online on June 1st, while the publicly reported vacancies amounted to only 151 units on average (see Table 5.a).

Year	Number of vacant units on June 1st	Number of listings online on June 1st	Number of listings online during whole year	Number of listings in % of stock
2004	307	891	16,293	8.1%
2005	151	1408	15,101	7.5%
2006	259	1394	15,247	7.5%
2007	180	1244	26,729	13.0%
2008	57	709	13,112	6.3%
2009	109	599	13,977	6.7%
2010	136	1248	14,675	7.1%
2011	125	367	5,844	2.8%
2012	206	124	3,658	1.7%
2013	242	480	4,852	2.3%
2014	471	759	6,613	3.1%
2015	483	1586	15,582	7.1%
2016	484	2930	13,930	6.3%
2017	454	3402	24,961	11.2%
2018	447	4776	26,476	11.8%
2019	306	2643	24,914	11.0%
2020	339	2970	27,891	12.2%
2021	381	3345	31,937	13.9%
2022	161	2233	14,938[5]	6.5%
Average (2004-22)	**279**	**1,743**	**16,670**	**7.7%**

Table 5.a Number of listings vs. number of vacant units (Federal Statistical Office, 2022)

[5] The reporting day of the dataset is December 23rd, why not the whole year is considered

At first glance, this may seem inconsistent, but upon further reflection, it makes sense. After all, the vacancy refers to all apartments that are effectively vacant on the reference date (June 1st). In a highly sought-after housing market like Zurich City, it is logical that apartments are immediately rented out again, ensuring there is no gap in the lease between the outgoing and incoming tenants. This means the apartment does not remain vacant in the interim and hence does not get accounted for in the vacancy statistics. Thus, based on this comparison, no statement can be made about how many of the apartments entering the market are represented in the dataset.

Therefore, it makes sense to revert to the approach of Chen and Biljecki (2022), who, in their study, compared the listings crawled during a three-month period from the web to the total housing stock and not only the vacant units. This was done to estimate what percentage of the total inventory is covered in their study and resulted in roughly 1/10 of the stock (ibd.).

In this thesis the approach is slightly adapted, as not a three-month period is available but several years of listings and several years of statistics on the stock (Stadt Zürich Statistik & GWZ, 2023). For this, not only the listings that were online on June 1st are considered, but all that were online during the reference year in question. This can be compared to the statistics of the housing inventory, which is published annually.

However, this approach does not rule out the possibility of duplicates. As described in Section 3.1, duplicates are grouped by the Meta-Sys to consolidate the same entry for different platforms and re-uploads within short times, but the following duplicates remain:

1. Listings that span a period starting in year X and ending in year Y are counted for both years.
2. Listings for the same apartment can appear multiple times in the dataset if they have been advertised several times over the long observation period (2004-2022).

Despite certain redundancies, this approach should be able to provide an estimate of what percentage of the stock is represented each year. Chen and Biljecki (2022) did not face this issue to the same extent, as their crawling horizon spanned only 3 months. As a result, the number of instances where the same apartment appears multiple times in the dataset would likely be significantly lower.

The results for this thesis are depicted in Table 5.a. Upon examining the table, it is noticeable that the proportion of advertised apartments in the dataset compared to the total inventory varies considerably. For certain years, a large number of listings can be detected, accounting for up to 13.9% of the stock. In other years, the proportion is much lower at 1.7%. On average, the listings represent 7.7% of the stock. This, of course, means that the coverage is significantly worse than in the study by Chen and Biljecki of the Singaporean market. There, with listings crawled over a period of 3 months, approximately 10% of the housing stock could be covered. In contrast, here, over a period that's four times longer (1 year), the average coverage is only 7.7%.

5.1.2 Comparison with movements

As already introduced in section 2.4.4, in Switzerland, every relocation process is accompanied by a reporting procedure resulting in two relevant statistics: Number of incoming and outgoing moves to and from the city, and the number of relocations within the city (Stadt Zürich Statistik, 2023f, 2023d).

As derived in section 4.1.1, the combination of incoming and relocating residents, when adjusted for household size, should provide insight into how many apartments are newly occupied and thus could have theoretically been publicly listed on the market. It's important to note that statistics regarding the average household size are unfortunately only available for the years 2013 to 2021. Data for 2022 is yet to be published, and earlier years were not publicly accessible. However, given that the household size barely changed noticeably within the observed span (ranging from 1.97 to 2.0 persons), the 2021 household size was assumed for 2022, and the 2013 household size was used for the earlier years from 2004 to 2012.

The subsequent Table 5.b presents the individual metrics and their adjustments, comparing them to the number of listings.

Year	Number of persons, incoming to and relocating within Zurich City	Resulting newly occupied apartments in Zurich City	Coverage of newly occupied apartments through listings in the dataset
2004	77,438	39,309	41.4%
2005	79,792	40,405	37.3%
2006	80,602	40,915	37.3%
2007	88,659	45,005	59.4%
2008	80,449	40,837	32.1%
2009	80,899	41,065	34.0%
2010	81,046	41,140	35.7%
2011	83,440	42,355	13.8%

2012	81,880	41,653	8.8%
2013	90,952	46,169	10.5%
2014	85,217	42,823	15.4%
2015	87,821	44,131	35.3%
2016	87,889	44,165	31.5%
2017	90,983	45,492	54.9%
2018	89,933	44,967	58.9%
2019	88,392	44,418	56.1%
2020	82,742	41,789	66.7%
2021	84,940	42,899	74.4%
2022	89,197	45,049	33.2%
Average	**84,856**	**42,873**	**38.8%**

Table 5.b Number of listings in the dataset compared with moving statistics, calculations based on (Federal Statistical Office et al., 2022; Stadt Zürich Statistik, 2023f, 2023d)

Given that the household size is nearly an average of 2 persons per household, the estimated number of newly occupied apartments is roughly half of the number of relocating and incoming individuals. During the observed period, an average of 42,873 apartments per year were newly occupied. When compared to the published listings, there are significant fluctuations, mirroring the varying entries in the dataset. In the dataset, the number of listings ranges from 8.8% to 74.4% of the total number of newly occupied apartments. Nevertheless, it's evident that the number of newly occupied apartments significantly surpasses the number of listings, confirming the "off-market" leasing of many units. Using this approximation method, on average, 38.8% of the apartments are not listed and therefore not openly available on the free market. Conversely, more than half, or 61.2%, of all apartments are informally transferred, for example, when the outgoing tenant directly suggests a successor. Naturally, this reduces effort for the property managers, as they don't need to handle viewings or sift through hundreds of applicants (see also section 2.4.4).

5.1.3 Correlation between description lengths and vacancy rate

After confirming the influence of the highly-competitive residential market in terms of the reduced number of listed apartments, we now turn to the question of its impact on the remaining listings. As described in section 4.1.2, the average number of words per description is determined, as this is intended to serve as a simple approximation for the comprehensiveness of a listing – the more words, the more detailed the description; the fewer words, the poorer the description.

The average word count for the listings in the studied municipalities could be extracted using R. On average, over the entire time span covered by the dataset, an apartment listing in Zurich City contains 92 words (see Table 5.c). This is slightly below the average of 95 words for all considered municipalities (Zurich and neighboring municipalities). At the same time, Zurich City, with a vacancy rate of 0.13%, is significantly below the average of 1.0%.

Municipality	**Average number of words per description**	**Average Availability rate (in % of stock)**
Adliswil	85	0.85
Dübendorf	86	0.86
Fällanden	99	1.75
Kilchberg	109	0.89
Maur	106	1.00
Oberengstringen	90	0.75
Opfikon	89	1.35
Regensdorf	92	0.77
Rümlang	81	1.35
Schlieren	82	0.28
Stallikon	117	1.26
Uitikon	112	1.30
Urdorf	83	1.01
Wallisellen	94	1.36
Zollikon	103	0.73
Zurich City	**92**	**0.13**
Average	**95**	**1.0**

Table 5.c Average number of words per description and municipality

Calculating the Pearson correlation between the number of words and the availability rate, it amounts to 0.14. The average availability rate and the average number of words per description have therefore a very small positive correlation. Based on the chosen approach, one would have to conclude that the influence of the vacancy rate on the quality of the listing descriptions is virtually non-existent.

At the same time, it must be pointed out that all the markets under consideration are characterized by low vacancy rates. All of them have vacancy figures below what would indicate a balanced market, which, according to experts, stands at a vacancy rate of 1.5% (Thalmann P, 2012). This is unsurprising, as the market pressure in the city is known to also impact the surrounding municipalities. For the matter at hand, this implies that an examination across more municipalities, especially those with higher vacancy rates, might have potentially led to different results.

5.2 Pre-Processing of data

The filtering of relevant data and identification of doubles leads to a loss of data of roughly 27% in both cases (Zurich and Suburbs and Zurich City only). There are no major differences for the different processing steps, regardless of whether only Zurich City or Zurich City and suburbs are processed. Both selected geographies show very similar relative numbers for listings with no title and text description, listings with imprecise geocoding and self-identified doubles with same address and text information (see table 5.a).

Data	**Zurich City only**	**Zurich City + neighboring communities**
Number of entries total	274,561 (100%)	421,602 (100%)
Number and share of entries with empty information in title and description	368 (0.13%)	557 (0.13%)
Number and share of entries with imprecise geocoding	63,587 (23.16%)	99,540 (23.61%)
Number and share of entries after exclusion of (self-determined) doubles	11,108 (4.05%)	15,894 (3.77%)
Number of entries remaining after all exclusions	**199,498 (72.66%)**	**305,611 (72.49%)**

Table 5.d Overview on number of entries for different steps of data pre-processing

The remaining listings to be analyzed for Zurich City refer consequently to 199,498 listings.

5.3 Named Entity recognition with spaCy

Following the quantification of the data and subsequent filtering and deduplication of duplicates, we now proceed with the Named Entity Recognition (NER) of spaCy for the "Location" type.

5.3.1 Results of initial model without improvements

In the initial run of Entity Recognition, many location references were tagged as "Locations" by spaCy, but they are not actual locations or places in the sense that would be useful for this analysis (see Figure 5.1).

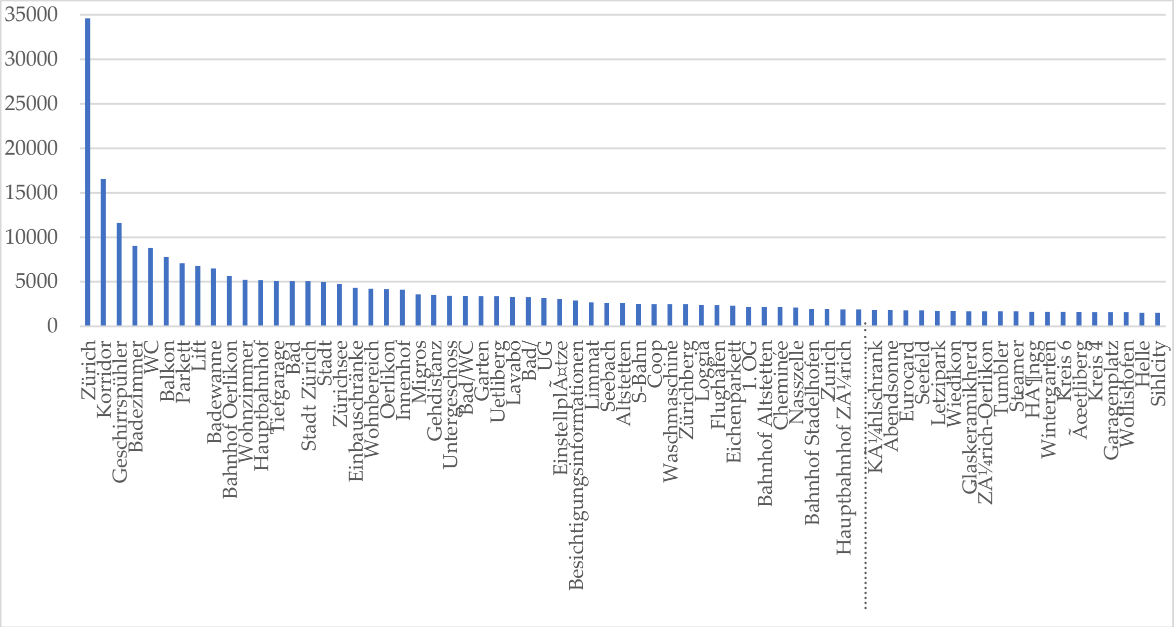


Figure 5.1 Recognized locations by spaCy number of ocurring for locations occurring more than 1'500 times

This includes location references within the housing unit itself. Classic examples of this include the bathroom and the corridor, which are among the top 10 most frequently occurring "Locations." These appear frequently in the listing descriptions when describing the interior and were detected by spaCy as "Locations." This is not incorrect, as of course, these are locations, but for the purposes of identifying place names related to the locality of the apartment, it is not suitable. For this reason, these are also counted as incorrectly detected toponyms for the initial calculation of precision and recall.

The most frequently found toponym by far is "Zurich" (found 34,606 times) which is unsurprising, as this location reference applies to all listings. However, this also makes it impossible to conduct an interesting analysis with it, as the term appears throughout the entire analyzed dataset and geographical area. However, since this term appears by far the most frequently and would strongly distort the weighted calculations, even though it is not relevant for the further analysis, the term is completely removed, thus not affecting the counts of both falsely and correctly detected toponyms.

In this context, it also becomes evident that what is clearly recognized in everyday language as the same location reference is counted by SpaCy as different location references because the words don't exactly match. For instance, the terms "Stadt Zürich" (City of Zurich) and just "Stadt" (City) are among the most frequently detected locations, both of which also correspond to Zurich. Accordingly, they are treated the same way as the top candidate, Zurich (removed and not included for weighted calculations of Precision and Recall).

Furthermore, among the frequently detected toponyms, location references can also be found that are clearly located outside the city. This includes, for example, the airport, which is located in the municipality of Kloten. It is often mentioned as a selling point because of its proximity, but it is not an immediate location reference for the actual micro location of the apartment. In the listing descriptions, you may find phrases like "close to the airport" or descriptions of the accessibility of the airport, which is easily and frequently connected to Zurich by tram and train. The same applies to other major Swiss cities, which, although not among the top localities according to the following figure, were also frequently detected.

The last, incorrect, frequently occurring, and somewhat peculiar detection was a sequence of dots ("..........."), which spaCy extracted as a "Location" 1,874 times. This sequence often appears in the advertisement texts as a stylistic device to separate different sections, but of course, it has no relation to an actual location.

The manual control of a sample showed various terms that spaCy did not recognize although of relevance for the analysis. In comparison to the detected ones, these unrecognized location references largely don't appear to be different, and determining why some were found and others weren't seems challenging. Notably, various descriptions associated with the term "location" (e.g., central location, quiet location, convenient transportation location) were not detected. As a result, these will be addressed using a n-gram extraction (see section 4.3.4). The other terms, which might be descriptions like "in the heart of Zurich" or "in the midst of the city of Zurich," or simple street names ("Sophienstrasse") or neighborhoods ("Allenmoosquartier"), are added to the "look-up" list. This list is fed back into Python, and its terms, alongside the locations detected by spaCy, are also extracted.

5.3.2 Precision and Recall (before and after adding n-grams and improving model)

After removing the special case of "Zürich" and all its alternative spellings, a manual detection of the identified location names is conducted, categorizing them into "correctly identified" and "incorrectly identified." In this process, the first 3,025 locations, each having 17 or more entries, are reviewed. In total, there are 54,968 entries. The detected 3,025 entries thus constitute only 5.5% of all entries. However, when considering the number of detections per entry, these 3,025 comprise 8.1% of all data. This should be sufficiently representative for determining the Precision, especially the weighted Precision.

For the reviewed entries, the Precision can then be calculated as described in section 4.3.2. The Precision for this initial round was relatively low due to the many "incorrect" locations in both the weighted and unweighted cases (see Table 5.e) and refers to 30.8% for weighted precision and 36.7% of not weighted precision. This is because among both the frequently occurring and less frequent location references, there was a high number of inaccuracies.

The manual detection of 200 advertisements yielded a considerably better value for Recall (as shown in the table). This is not surprising, as poor Precision often leads to good Recall, given that many entries, even if incorrect, are identified. To determine this, 212 listings were manually checked, and the detected location references and toponyms were compared with the ones found. Among the 212 listings, 113 location references were identified, of which 84 were detected. In the initial round, the Recall is thus 69.7% (see Table 5.e) and it was determined without weighting since it was only assessed for a relatively small sample.

	First Round		Final Round	
	Precision	Recall	Precision	Recall
Weighted	30.8%	n/a	100%	n/a
Not weighted	36.7%	69.7%	100%	74.3%

Table 5.e Precision and Recall for first and final round of NER

Improving precision can be easily achieved: all incorrectly detected location references are removed, resulting in automatic precision of 100% for the final round. To enhance recall, n-grams around the term "Lage" (location) are added, along with the other location references not found during manual review. In the end, with these improvements, the precision reaches 74.3% for another sample of 200 listings, which is deemed sufficient.

5.3.3 Extracted location references

After manually consolidating the ambiguities and adopting the most-referenced toponym (such as assuming "See" refers to Lake Zurich), a final list is established containing all location references with 200 entries or more. This list comprises exactly 125 location references, of which the most frequent ("zentrale Lage" or "central location") appears 25,943 times, followed by "ruhige Lage" (quiet location) with 14,597 entries, and Zurich Main Station (often referred to in the original simply by the abbreviation "HB" instead of "Hauptbahnhof") with 9,514 entries (see Table 5.f).

Location Reference	**Number of Counts**
zentrale Lage	25943
ruhige Lage	14597
Hauptbahnhof Zürich	9514
Zürichsee	6318
Bahnhof Oerlikon	6206
Oerlikon	6181
Uetliberg	5652
Seebach	4256
Altstetten	3933
Höngg	2854
Wollishofen	2798
Zürichberg	2787
Limmat	2680
Seefeld	2425
Witikon	2294
Bahnhof Altstetten	2286
Wiedikon	2252
Bahnhof Stadelhofen	1916
Affoltern	1825
Kreis 6	1789

Table 5.f The top 20 location references in descending order

The complete list of toponyms with more than 200 entries can now be categorized, within which the spatial results of the KDE can subsequently be discussed. The subsequent chart (see Figure 5.2) illustrates the distribution of the 125 place references across the various defined categories, which will be elaborated upon in detail and exemplified in the sections that follow.

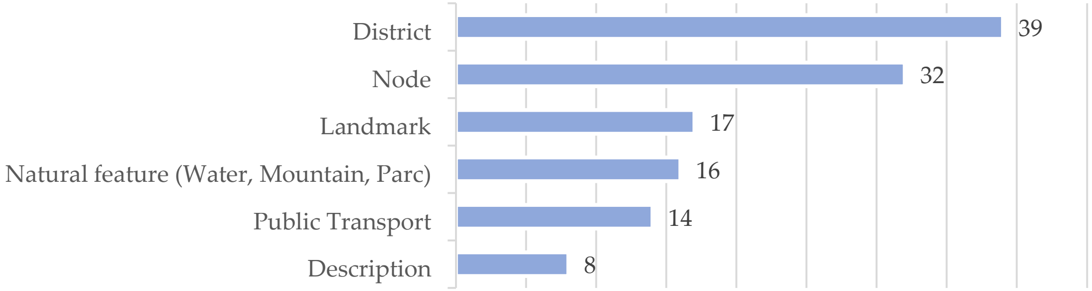


Figure 5.2 Number of Location References (> 200 entries) distributed among categories

The category that appears by far the most frequently is the "District" category with 39 of all location references being a district. Inspired by the districts described by Lynch (1960), this category encompasses all designations that point to an urban area. This includes officially recognized city districts or city circles, such as Oerlikon, Altstetten, or Kreis 6. Additionally, there are vaguer districts in line with vernacular geography, such as "Innenstadt" (Inner City) or "Trendquartier" (Trendy District).

The subsequent category to be addressed is "Nodes", drawing inspiration from Lynch's urban analysis (Lynch, 1960). As the name suggests, this category contains references, in most cases, to squares or other places. This category includes actual squares, as we might typically imagine them, like Paradeplatz, Kreuzplatz, or Limmatplatz. It also encompasses other clearly identifiable places in the city, which aren't landmarks and aren't clearly associated with public transport, like Römerhof or Schmiede Wiedikon. In this sense the category shares a lot with Lynch’s category of knodes. The category also includes the detection that can specifiaclyl be assigned to public transport.

Included in the nodes is also "Public Transport" as for the train station they are extracted as “station XX” and therefore this node is clearly identifiable as public transport. Proximity to public transport is an important selling point in a property advertisement, so it's interesting to explore how widely the related stop is referenced. But because stops are often just extracted without context, it's frequently unclear whether a reference is to the stop itself or the associated place. For instance, while Schmiede Wiedikon is a significant transport junction for various bus and tram lines, Farbhof is also the terminal stop for the Tram 13 line, but still they could be places as well. Here, however, only the clearly detected public transport stops are included, specifically train stations like main station, station Oerlikon, station Binz, etc., thus excluding tram or bus stops.

The third category is "Landmarks", also inspired by Lynch's work (Lynch, 1960). These are easily recognizable building complexes that are particularly suitable for orientation. In the dataset at hand, examples include Letzipark, Sihlcity, the Opera House, Messe Zürich, or the Children's Hospital.

In a way, the "Body of Water", "Mountain", and "Park" categories can be termed as "natural landmarks". The first includes, for instance, Zürichsee or Katzensee; the Mountain category frequently references Uetliberg, and the Park category, Irchelpark.

The "Description" category encompasses descriptions of the location. This includes the already known location descriptions, namely "ruhige Lage" (quiet location) and "zentrale Lage" (central location), as well as "Grünen" (in greenery), "am Waldrand" (at the edge of the forest), or "Zentrumsnähe" (near the city center).

5.4 Spatial analysis through KDE

In the subsequent step, the 125 selected toponyms, each with 200 or more hits, can be spatially represented. This amounts to a total of 178,867 data points. An initial representation of the data points from all categories indicates that they are mainly distributed across the entire urban area of Zurich-City (official city boundary depicted in red) (see Figure 5.3).

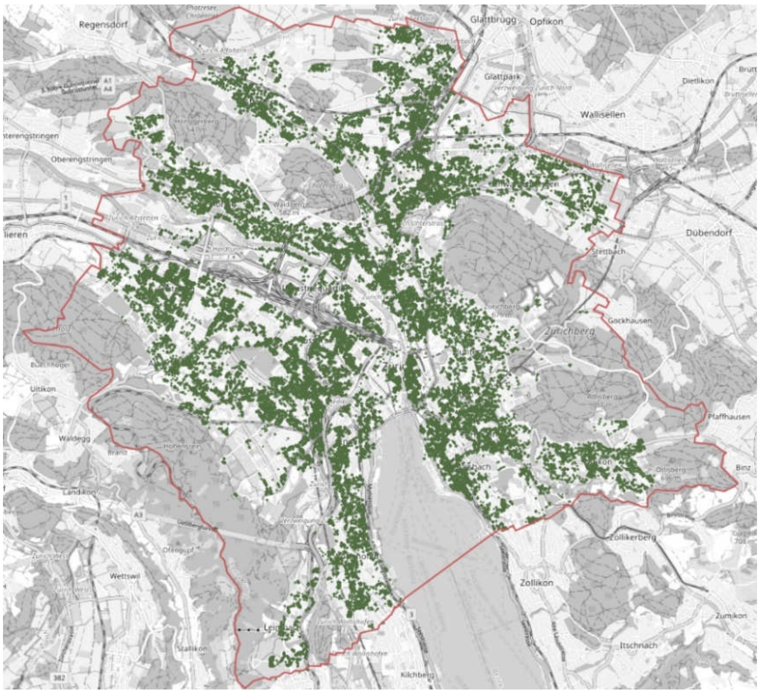


Figure 5.3 Distribution of location entities (ocurring more than 200 times)

An exception can be found at the quarter of Friesenberg, where not many datapoints occur, or not the whole area is covered. In the west, this can undoubtedly be attributed to the presence of large allotment garden areas, although they are not easily discernible on the map. However, even in the east, where the settlement area commences, there are fewer data points compared to the rest of the city. The Friesenberg district, in particular, does not appear among the examined toponyms.

In the remaining course of this chapter, the results of the KDE for selected place references from each category will be presented, described, and interpreted. For all KDE maps, the kernel size is set at 100 meters, and the background, for orientation and toponym resolution, comes from the VGI-project OpenStreetMap (OSM). Any polygons on official borders for the city itself or city quarters and districts as used within the "District" category for comparison purposes come from the City of Zurich.

5.4.1 KDE Category District

Official City quarters

For the official city districts and circles, a few examples are showcased and discussed, particularly in comparison with the official district boundaries. In all cases an initial comprehensive overview of the generated heatmap is provided to discern its extent. In some cases, for more detailed discussion, the view is zoomed in to better delineate the relevant perimeter.

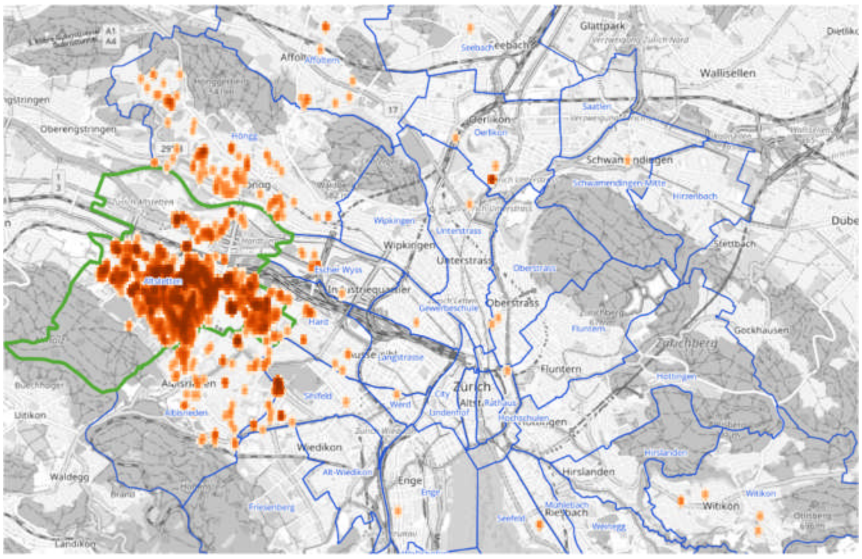


Figure 5.4 Overview on whole KDE of district Altstetten

The district of Altstetten (3,933 entries) was selected as a typical, expansive district with frequent occurrence and a high degree of recognition.

Individual kernels are scattered far and wide across the entire city. However, the concentration is most pronounced in and around Altstetten (in green), including the districts of Albisrieden and Höngg.

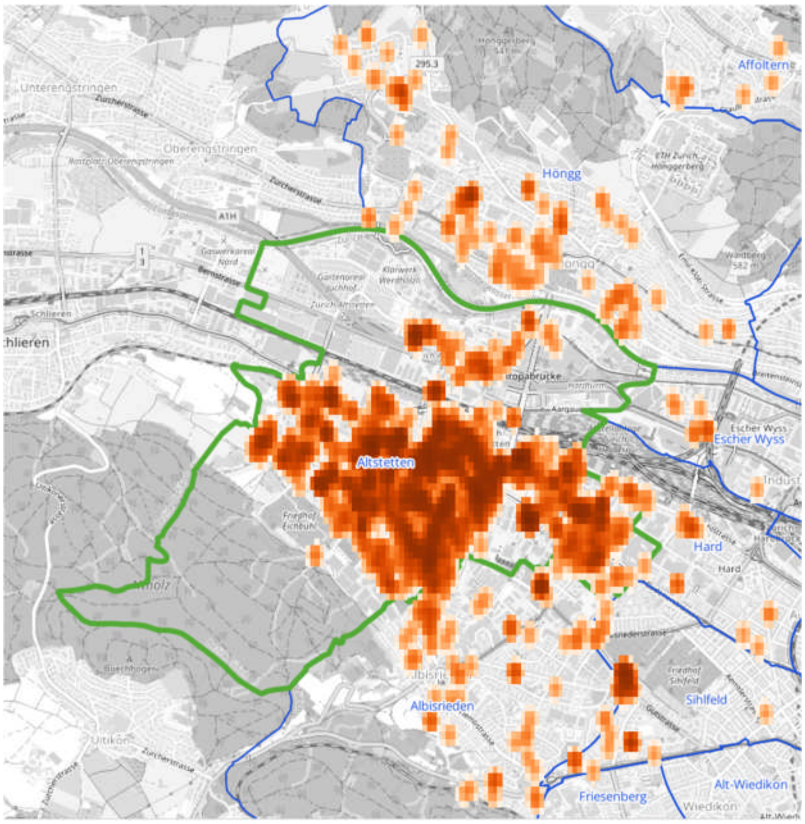


Figure 5.5 District of Altstetten, zoom-in

The toponym resolution for the Altstetten district reveals that most data points are indeed located within the district. However, there's a notable lack of a sharp delineation towards the neighboring district of Albisrieden. This could be attributed to the fact that Altstetten, in contrast to the relatively quieter district of Albisrieden, is more well-known and more easily identifiable for many. Additionally, the exact boundary between the two districts is somewhat ambiguous. For instance, between districts 3 and 4, the Badenerstrasse serves as the boundary, and given its significance and prominence, it would likely be classified as a "edge" in the context of Lynch's (1960) terminology, what makes it easy to distinguish district 3 and 4. Conversely, the boundary between Altstetten and Albisrieden doesn't follow a prominent road but rather weaves through the district, which makes it more difficult to assign a location close to the border to one of the two districts. Many data points are also found in the elevated district of Höngg, which might be attributed to descriptions emphasizing the view over Altstetten, given its sloped positioning.

For Seefeld (2,425 datapoints), Analogous to Altstetten, kernels are distributed across the entire city here as well, though it's unclear why they appear in such distant locations. For the most part, the vernacular Seefeld is concentrated on the right lakeshore, with its presence diminishing as one moves further away from it.

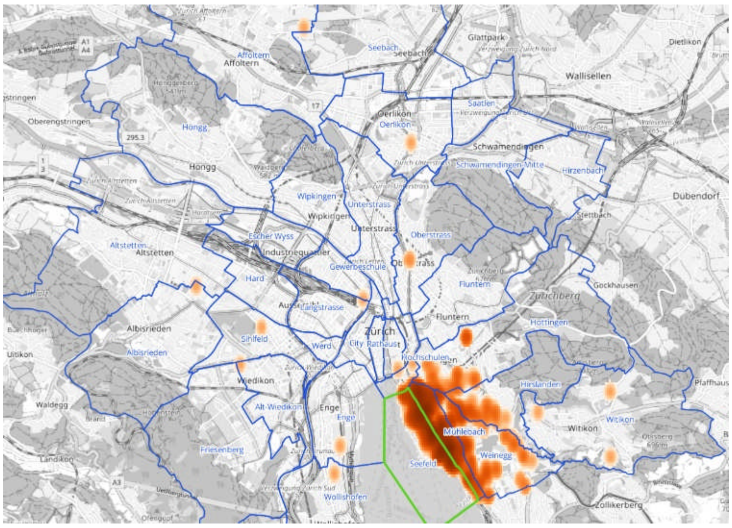


Figure 5.6 Overview on whole KDE of district Seefeld

Upon closer examination, it becomes unmistakably evident that Seefeld, with its high density, also fully encompasses the neighboring district of "Mühlebach" and even extends beyond it. It covers the populated parts of Weinegg as well as the adjoining western parts of Hottingen and Hirslanden. All of these appear far less frequently in the dataset and do not fall under the analyzed location references/ toponyms with 200 or more entries. This suggests that Seefeld, being a prestigious and well-known district located directly by the lake and near the city center, is either marketed beyond its boundaries as a selling point or is perceived as such because it's more recognized than the other quarters in District 8. According to the official boundary, Seefeld extends to the middle of Lake Zurich. Naturally, no data points are found there, given the absence of any settlement area.

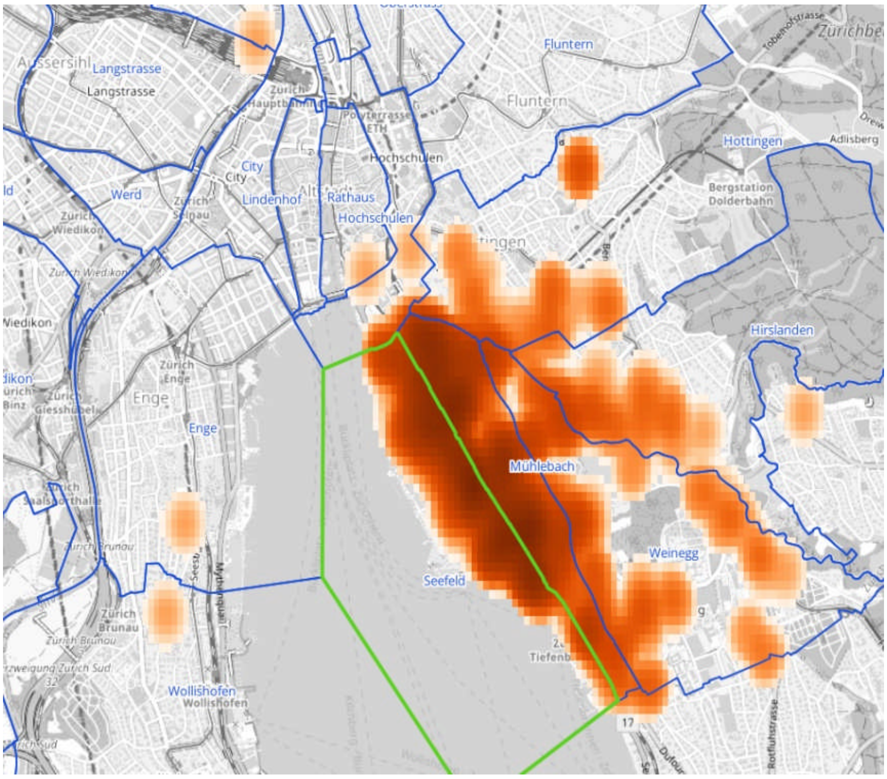


Figure 5.7 KDE of district Seefeld, zoom-in

Whether the lake would also be perceived as part of Seefeld and not just as a lake is probable. However, this question cannot be conclusively answered based on the current analysis.

Non-official districts

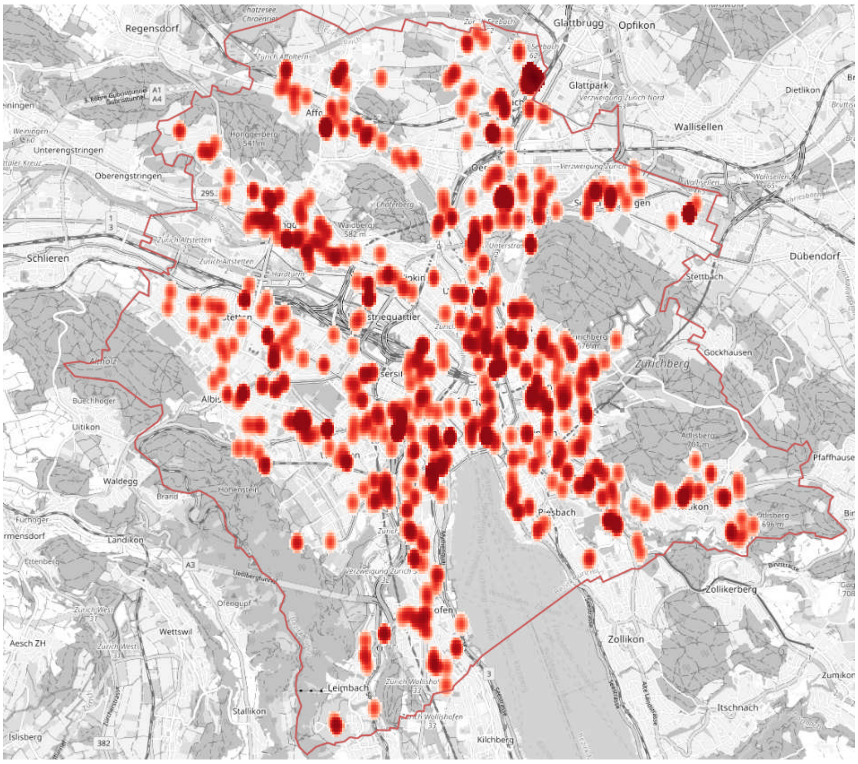


Figure 5.8 KDE of district "Innenstadt"

The term "Innenstadt" (inner city, 1,469 entries) exhibits virtually no spatial pattern, being mentioned uniformly across all parts of the city (city border marked as red line). In this regard, the current analysis is not suitable for capturing the vernacular geography of the inner city. A quick review of individual examples sheds light on the reason. In some instances, the term "Innenstadt" is likely used to describe a specific location. In many other, more peripheral cases, the "proximity to the Innenstadt" or the quick accessibility to the inner city due to a good public transport connection is highlighted. Thus, according to Zurich real estate listings, virtually the entirety of Zurich is either close to or within a short travel distance to the inner city.

In contrast to the term "Innenstadt", the vague descriptor "Trendquartier" (trendy district, 370 entries) is not ubiquitously used across the city. Instead, it seems reserved for areas genuinely considered to be trendy. For better orientation and description, the official boundaries of the 12 city districts are additionally displayed alongside the heatmap. It becomes evident that the highest concentration of "trendiness" is perceived in Districts 4 and 5, as well as in the northern part of District 3 (roughly corresponding to the Sihlfeld neighborhood).

This observation aligns with expectations. As previously elaborated upon in section X, especially District 5 has undergone significant transformation over the past few decades, with numerous former industrial buildings repurposed for residential or other uses. This district has

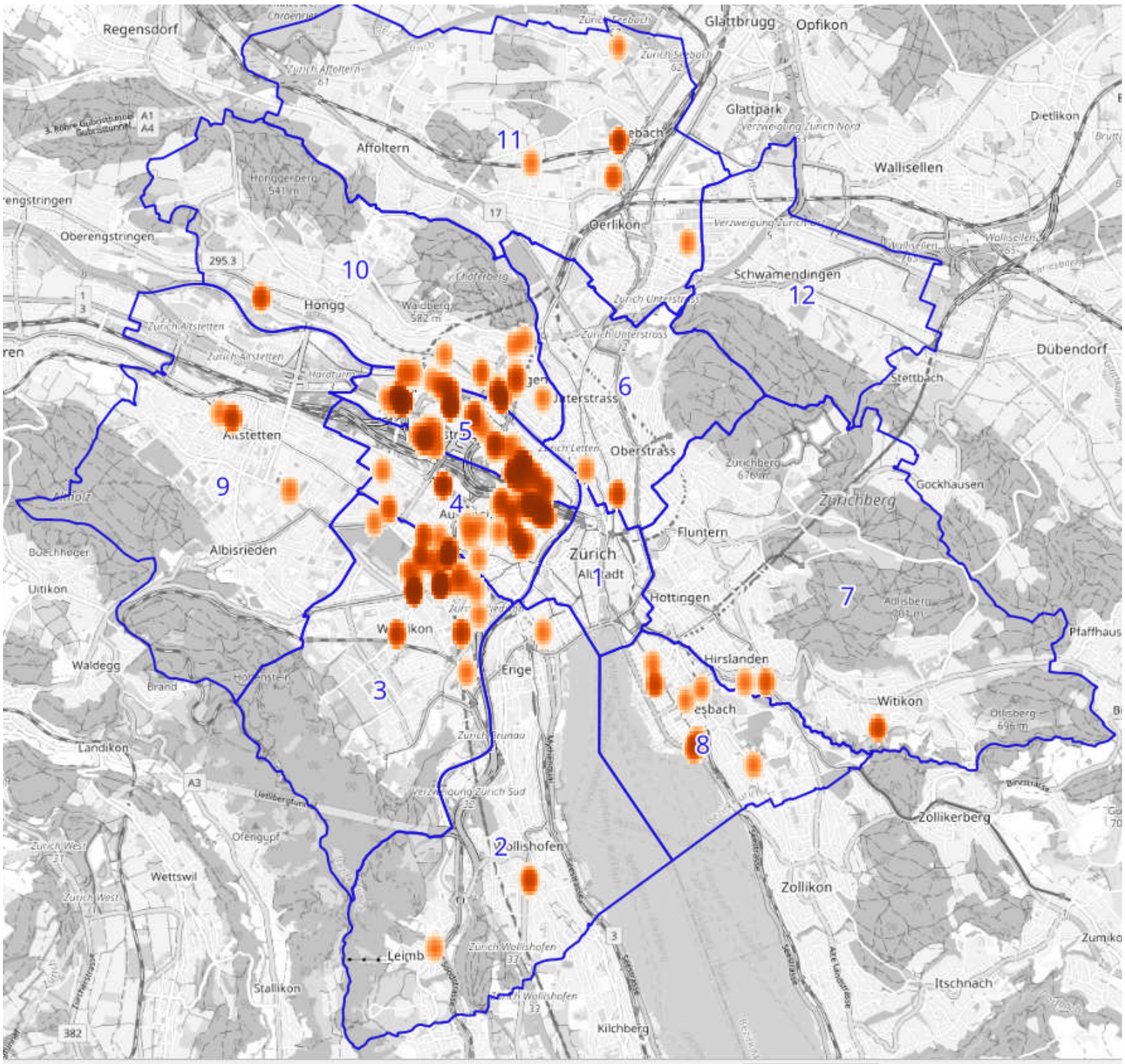


Figure 5.9 KDE of trendy district

evolved into a diverse, popular, and centrally located area (extending up to Zurich's main train station), boasting a myriad of restaurants, nightclubs, and some cultural venues established within the former industrial structures, such as the Maaghalle and Schiffbau. North of District 5, the Wipkingen neighborhood in District 10 is also perceived as trendy.

District 4 encompasses the Langstrassenquartier, where the majority of the kernels are located. This district is a popular nightlife area brimming with bars, restaurants, and shops, contributing to its trendy image. It has also been frequently labeled as a "Trendquartier" in other sources (e.g. Fuchs, 2008).

District 3, particularly in its perceived trendy portion, now hosts an increasing number of bars, restaurants, and a few clubs. However, its southern, less-central, and quieter section is not considered part of the trendy district. Trendiness seems to wane westward, ceasing roughly at the Sihlfeld cemetery's level, except for a couple of isolated hotspots around Albisriederplatz, a major transportation junction.

There are other isolated kernels as well, such as around the Altstetten train station – a district undergoing rapid development, which a few decades ago had a less favorable reputation − or in the city's northern part in Oerlikon, another area witnessing significant growth and serving as the central hub of the city's north.

5.4.2 KDE Category Nodes

The 15 most frequently occurring toponyms for the category of places, listed in descending order, are as follows: Paradeplatz (1,491), Kreuzplatz (1,178), Limmatplatz (1,160), Goldbrunnenplatz (975), Farbhof (778), Meierhofplatz (765), Römerhof (760), Schaffhauserplatz (756), Berninaplatz (728), Schwamendingerplatz (690), Stauffacher (670), Klusplatz (650), Albisriederplatz (630), Bellevue (627) and Bucheggplatz (549).

Here again, a similar phenomenon to that observed with "Innenstadt" appears to be evident. A Location as "Paradeplatz", which is the most mentioned place, is situated in District 1, which is predominantly occupied by commercial spaces. This implies that there is very limited residential space available in that area, and hence, there shouldn't be many property listings there. Nonetheless, there seems to be a frequent reference to these two renowned and prestigious locations in Zurich's city center.

Yet, even with places that are more relevant on a finer geographical scale, like a city district, rather than the entire city, listings appear within a broader vicinity of the actual site. It seems that the more prominent the location, the larger the surrounding area in which it is mentioned in a listing text. The significance of a place appears to be influenced by the presence of other prominent landmarks or points of interest nearby. If another feature of greater prominence is in close proximity, references to a particular location do not spread as widely, as the more notable feature tends to be referenced instead, as it could be the case for Bucheggplatz (see Bucheggplatz below).

The subsequent examples display one site that appear in a smaller region (Goldbrunnenplatz), one in a medium-scale area (Bucheggplatz) and one site that is cited in a more extensive area (Paradeplatz).

Goldbrunnenplatz (975 entries)

The listings mentioning Goldbrunnenplatz (975 entries) are primarily located at or around the Goldbrunnenplatz (marked as a black dot). However, even here, there are some outliers located somewhat farther away. In these instances, the proximity to the square itself often doesn't play a role, but rather the connection to public transportation. For example, in the farthest kernel group to the west, there's a reference to a bus route that leads directly to Goldbrunnenplatz. Serving various tram and bus lines, the Goldbrunnenplatz is also an important local transportation hub. Nevertheless, the map effectively visualizes the area in which references to Goldbrunnenplatz are still made, indicating that it must be one of the most significant squares in the neighborhood.

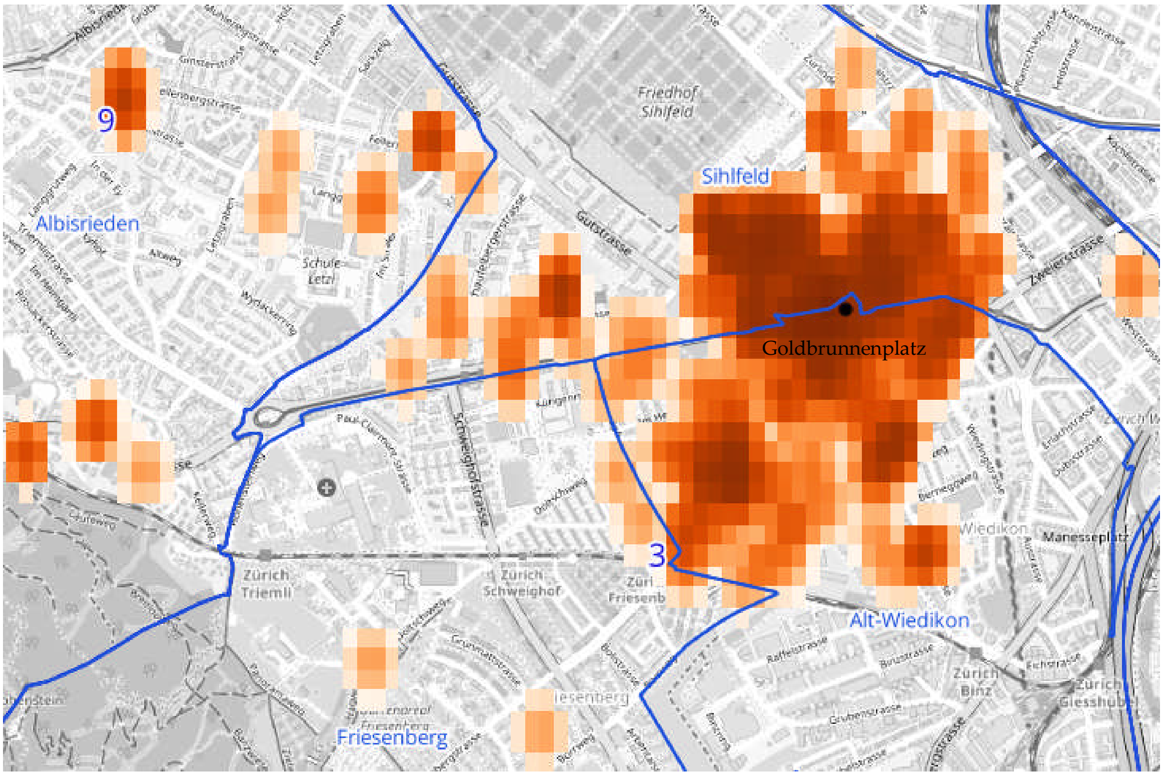


Figure 5.10 KDE of node of Goldbrunnenplatz (marked as black dot)

Bucheggplatz (549 entries)

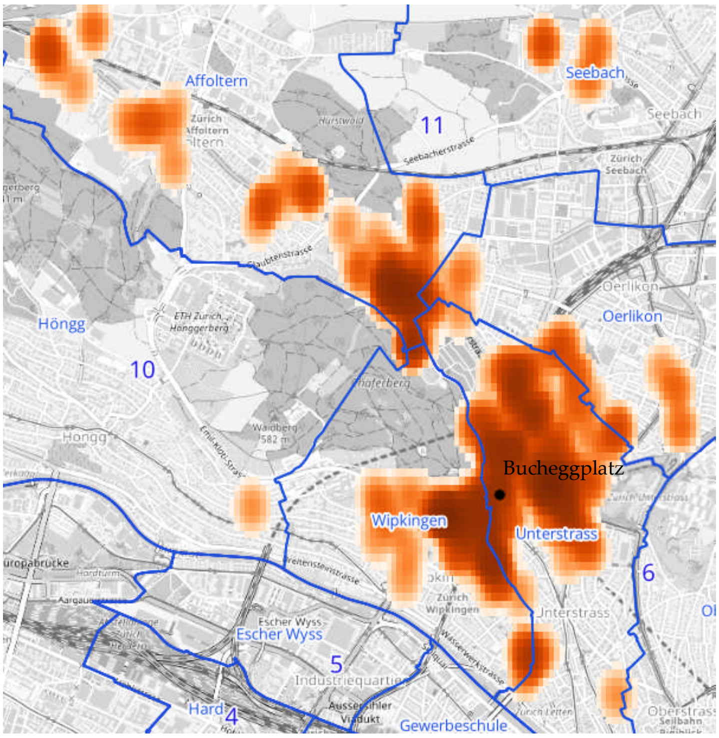


Figure 5.11 KDE of node Bucheggplatz

Mentions of Bucheggplatz predominantly extend into the Affoltern district, as it serves as an important junction for that area. However, in the Oerlikon district, despite a similar distance, references diminish more rapidly, potentially due to the presence and significance of Oerlikon Station.

Paradeplatz (1,491 entries)

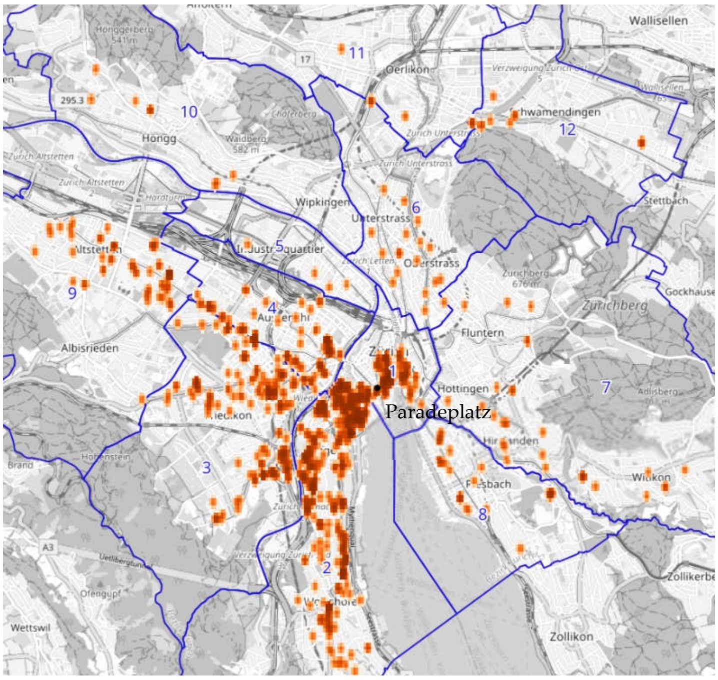


Figure 5.12 KDE of node Paradeplatz

As anticipated, the Paradeplatz is such a renowned location, arguably the most famous square in Zurich, that it is mentioned across the entire cityscape without a particularly strong spatial correlation. Intriguingly, there's even a gap directly at the Paradeplatz (indicated by the black dot). This is not surprising, considering the square predominantly consists of commercial buildings. For the Niederdorf area to the east and the Engequartier to the west, however, this square appears to hold significant relevance.

In the context of clearly assigned nodes of public transport, or in this case the referenced train stations, the trend observed is similar to that of general nodes. The larger and more significant the station, the broader the area in which it is mentioned. If there's another station nearby, mentions drop off rapidly. According to the advertisements, the five most significant train stations based on the number of mentions are the main station (“Hauptbahnhof”, 9,514), Oerlikon (6,206), Altstetten (2,286), Stadelhofen (1,916), and Wiedikon (1,432). The conspicuous absence of the similarly large Hardbrücke station among these top 5 might be attributed to its location between Oerlikon and main station, which quickly assume a more dominant role in mentions.

The following examples illustrate this trend for the most important and frequently cited station, Zurich Hauptbahnhof (Main Station), as well as the smaller Tiefenbrunnen station.

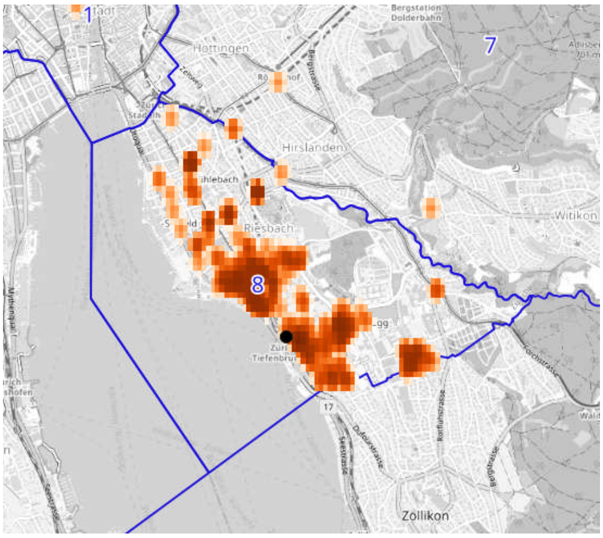


Figure 5.13 KDE of Public Transport Station Tiefenbrunnen

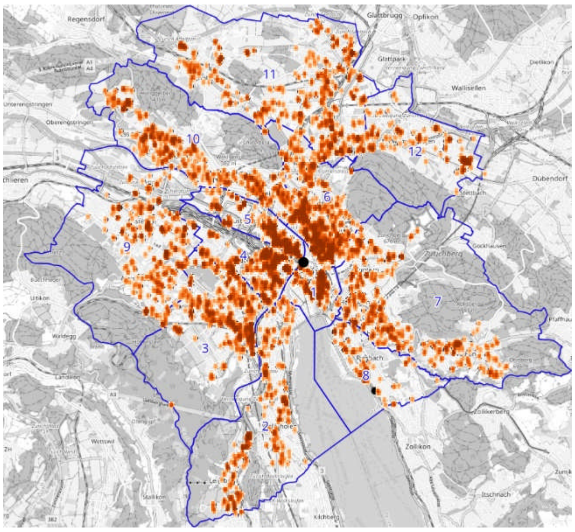


Figure 5.14 KDE of Public Transport Main station

5.4.3 KDE Category Landmarks

The 10 most frequently mentioned landmarks are Letzipark (1,732), Sihlcity (1,634), Triemli (1,350), Hallenbad Oerlikon (1,320), Universität (932), Irchel (907), Bad Allenmoos (726), Universitätsspital (622), Letzigrund (400), and Hallenstadion (345).

Interestingly, the list is dominated by two shopping centers at the top. Also significant are hospitals (Triemli and Universitätspital, often referred to with the abbreviation "Unispital"), public swimming halls (Oerlikon and Allenmoos), and stadiums (Letzigrund and Hallenstadion).

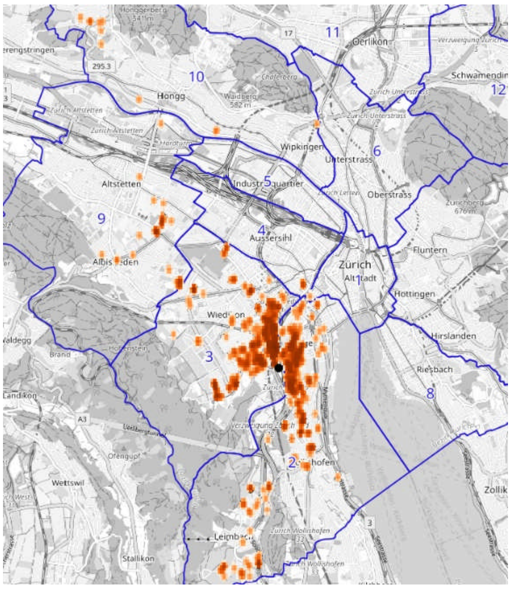


Figure 5.16 KDE of Landmark Sihlcity

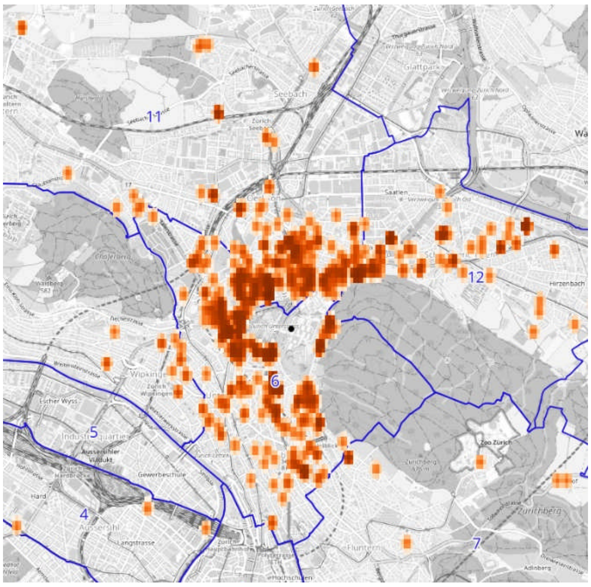


Figure 5.15 KDE of Landmark Irchel

It is intriguing that Irchel is mentioned so frequently, especially since there are no residential buildings directly adjacent to the university campus of Irchel. Surrounding Irchel is a vast park area, which explains why the advertisements are located at a considerable distance from the target landmark.

In contrast, for shopping centers (as illustrated by the example of Sihlcity in Figure 5.16), the listings are also indeed located nearby. Proximity to such a large shopping center is likely a significant selling point for everyday shopping needs.

5.4.4 KDE Category Natural Landmarks

Natural landmarks can be divided into body of waters, mountains and parcs and are mentioned for various reasons. They are either invoked in a general description extolling the virtues of the city of Zurich, as is often the case with the Lake of Zurich, or the Uetliberg and Käferberg as local recreational areas. Alternatively, they are highlighted for the vistas they offer, as is the case with mountains and lakes. The prime example, frequently associated with the view and with 1,271 entries also the most mentions mountains, are the Alps. While they are geographically distant from Zurich, on clear days their visibility can be quite prominent depending on one's vantage point. The third category of natural features are parks. In contrast to bodies of water and mountains, parks are referred to in a more localized context, often by describing the closest park to a given location.

Although the Alps are mentioned broadly across various parts of the city, a discernible pattern emerges that is logical. The majority of kernels are found along the elevated slopes of the Zürichberg to the east and the slopes of the Käferberg to the north. From these vantage points, the Alps, situated to the south of Zurich, are likely to be most prominently visible.

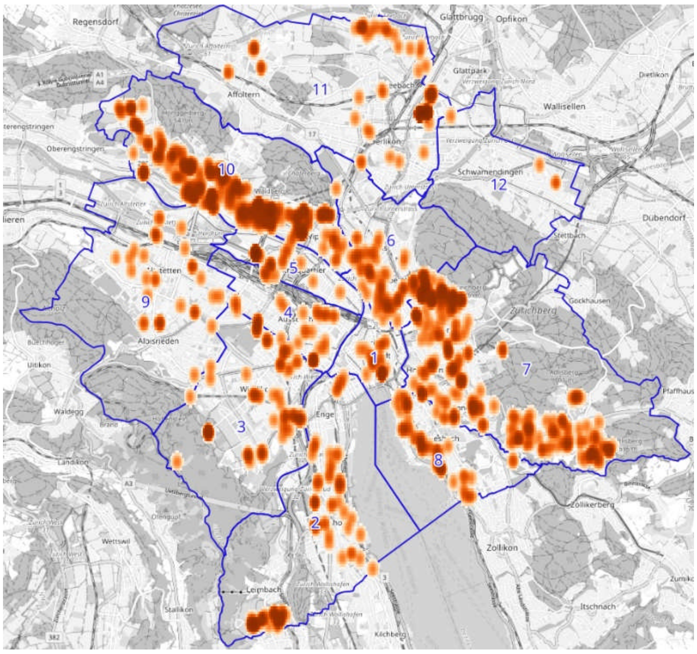


Figure 5.17 KDE of natural landmark Alps

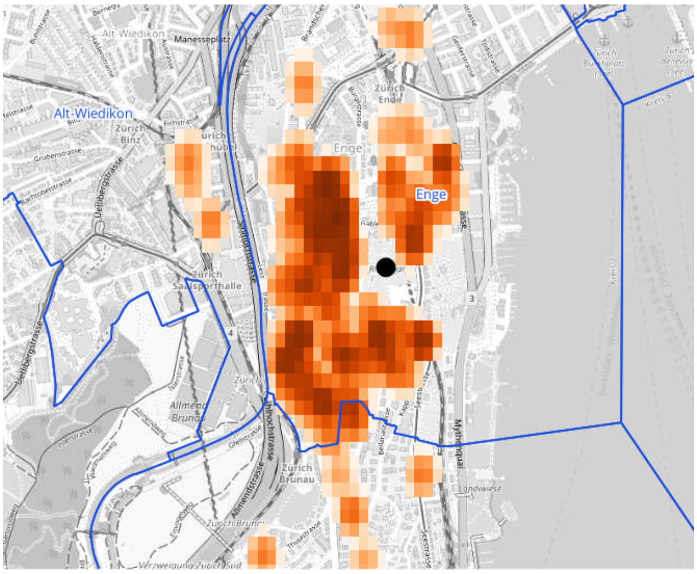


Figure 5.18 KDE of nautral landmark Rieterpark

The Rieterpark is a quintessential example of a park with local significance in the quarter of Enge. At the data point representing the park, there is naturally no mention, as the park area is not a residential zone.

Lake Zurich, with 6,318 entries the most frequently mentioned natural landmark, is referenced extensively in all habitable areas around and along the lake. Additionally, there seems to be a concentration of mentions along the hill slopes, as a lake view is an attractive feature. However, even in other areas, Lake Zurich is readily used as a selling point.

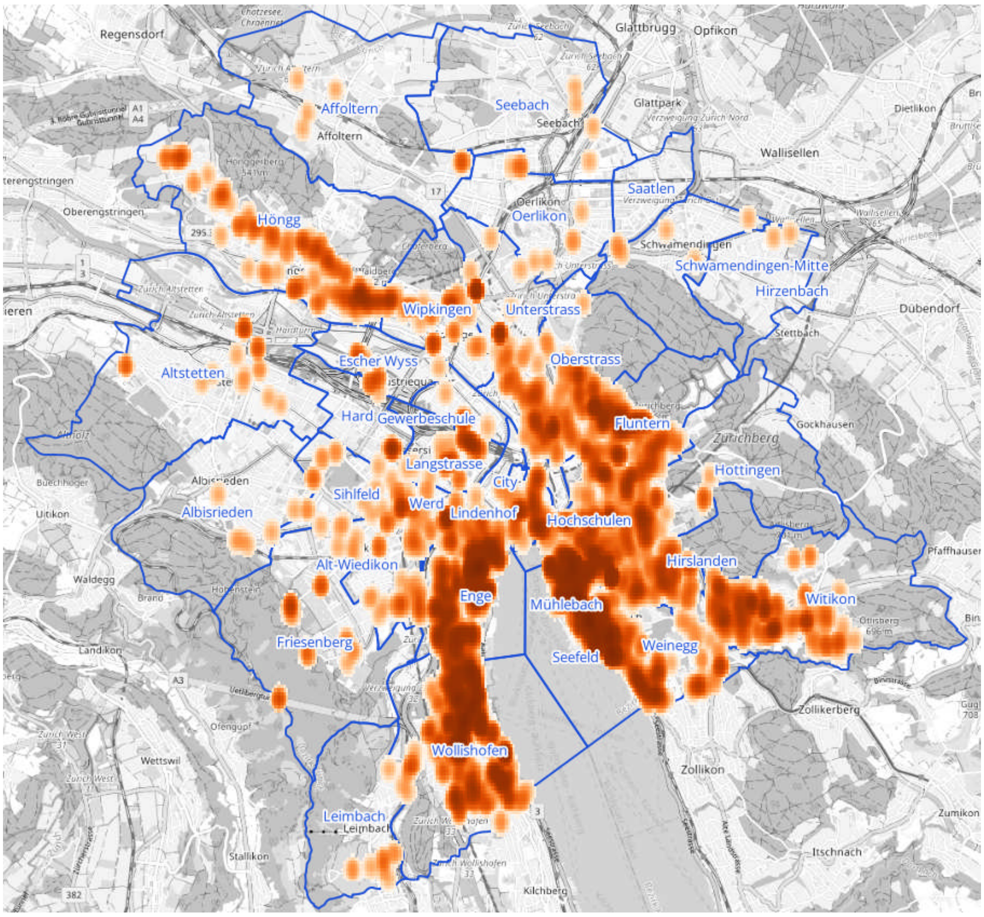


Figure 5.19 KDE of natural landmark Lake Zurich

5.4.5 KDE Category Descriptions

The descriptions of locations with more than 200 entries are as follows: central location ("zentrale Lage", 25,943), quiet location ("ruhige Lage", 14,597), greenery ("Grünen", 651, can appear as phrases like "quickly in the greenery"), prime residential location ("beste Wohnlage", 504), proximity to the center ("Zentrumsnähe", 300), and on the edge of the forest ("am Waldrand", 285).

For the frequently mentioned and generic descriptions like "central location" and "quiet location," they appear, as with all highly occurring locational references observed so far, to simply be selling points that are used ubiquitously. As illustrated by the example "quiet location" (see Figure 5.20) this descriptor can be found almost uniformly across the entire city, including in very central areas with nightlife that can't realistically be associated with a quiet setting.

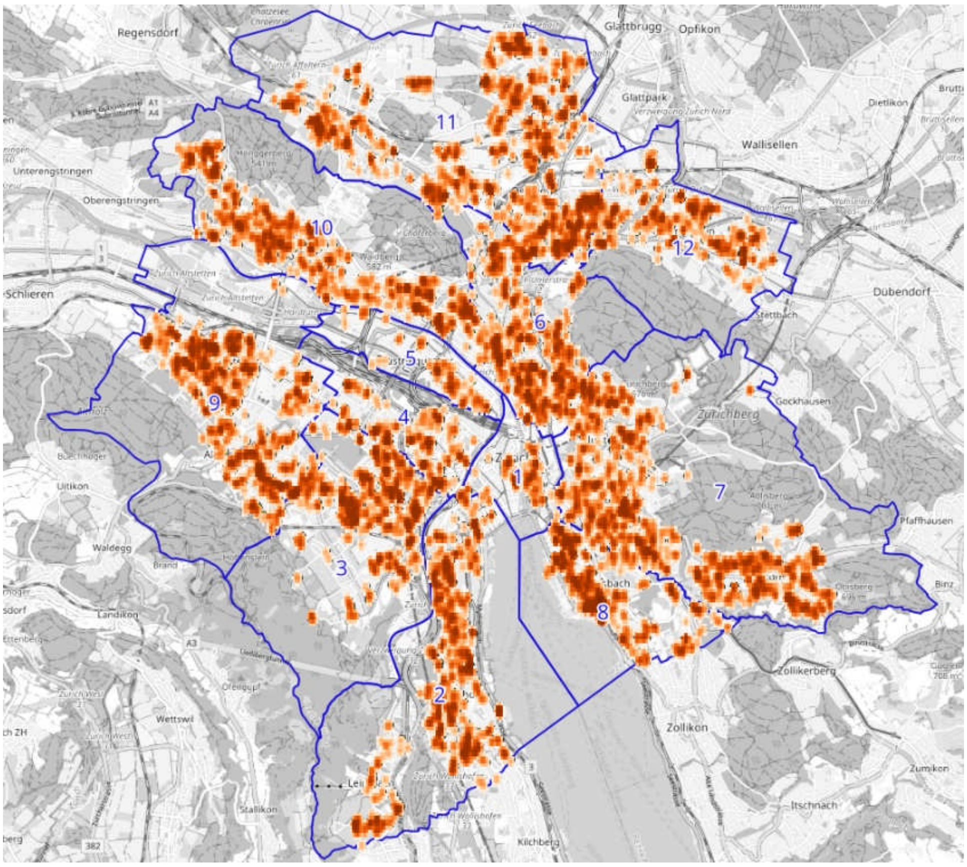


Figure 5.20 KDE of Description"quite location"

The less frequently used and more specific term "on the edge of the forest" exhibits a more logical pattern, and is indeed found almost exclusively along the forest edges in the western part of the city, along the forest edges of the Hönggerberg and Käferberg, as well as along the forest edges of the Zürichberg.

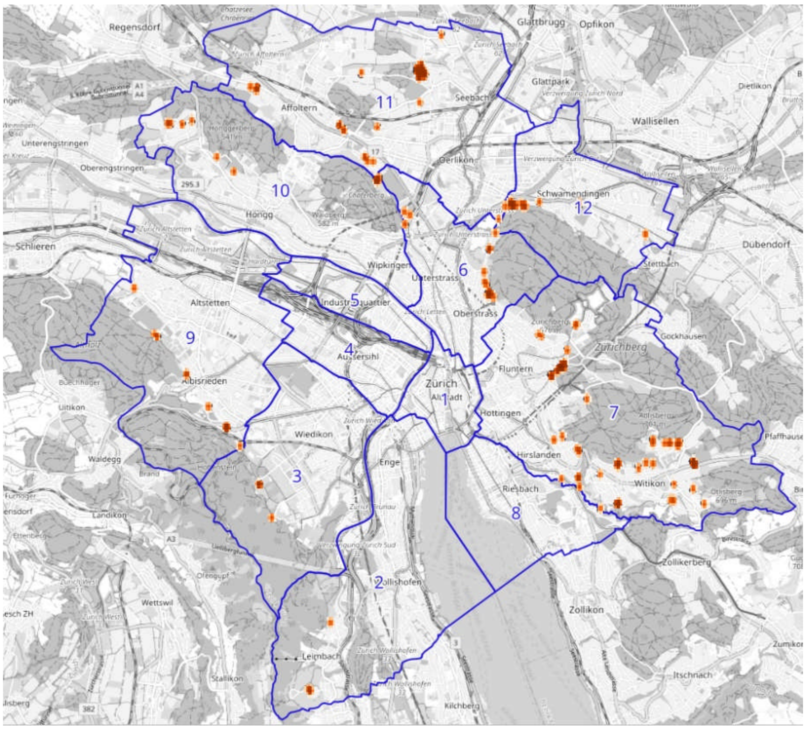


Figure 5.21 Description "on the edge of forest"

6 Discussion

This chapter focuses on the limitations of this study, the response to the initially posed research questions, and the significance of the insights derived from them.

6.1 Limitations

6.1.1 Dataset Scope

As corroborated by comparing relocation statistics with the dataset, the informal transfer of apartments in Zurich, due to the prevailing housing shortage (vacancy rate: 0.07%), is a significant issue. This comparison indicated that, on average, 61.2% of apartments never appear in online listings, which naturally limits the available real estate listings for evaluation.

In this context, the study also examined the extent to which the remaining listings might be influenced by market conditions. The hypothesis that this market environment could lead to "poorer" descriptions in the listings, given that property managers likely have a large number of applicants for nearly every apartment, and thus might be less inclined to write extensive texts, was considered.

This hypothesis was only inadequately substantiated using the chosen method (correlation between vacancy rates and the number of words in property descriptions). However, this raises the question of whether the chosen approach was sufficiently appropriate to conclusively address the question. On one hand, the comparison was only made between Zurich and its neighboring municipalities, all of which also have low vacancy rates. It would be interesting to re-examine the same approach for a broader set of municipalities that also exhibit high vacancy rates.

Moreover, there are more sophisticated methods in natural language processing to compare text descriptions than the mere count of words. Such methods include cosine similarity (especially tf-idf weighted word vectors), topic modeling, and perplexity. The cosine similarity is well-suited robust method for determining the similarity between texts (often represented as two vectors, frequently weighted by tf-idf) (Tata & Patel, 2007). Topic modeling refers to a statistical approach that employs various algorithms to investigate the semantic and thematic structure within a large collection of texts (Kherwa & Bansal, 2020), which would in this sense also be suited to investigate on the text and their differences in the different listings.

6.1.2 Extraction limited to “locations” without context of their mentioning

This issue was most evident with toponyms that should have primarily focused on a single point or a very specific location. Specifically, with the classes of nodes and public transport stops, which often produced widely dispersed density maps.

In this context, a significant limitation is the absence of extraction from entire expressions wherein toponyms are used. This led to these toponyms frequently appearing within a broad geographic perimeter. With a more refined differentiation, one could discern whether the respective housing unit is described as genuinely being "near" a certain place, or if, as seen in many more remote instances, a connection with public transport to that place is described, as is often the case for the terminus of a bus or tram line. In such scenarios, the referenced place may appear at a significant distance on the opposite side of the city, as in descriptions like "direct connection to the tram line 9 towards Heuried." In such cases, the resulting surface maps are probably better interpreted as areas where the extracted toponym has relevance, rather than fuzzy place indicating the geographic extent to which this place is recognized or by considering the given location as perceived to be near that place.

A similar issue arose, albeit to a lesser extent, with the category of districts. Here too, an extended extraction would have been beneficial to differentiate whether the housing unit is located within the respective district, is close to it, or is referenced in another context related to the district. For instance, in the elevated district of Höngg, the district of Altstetten is often mentioned simply because the view overlooking it is incorporated into the property listing description.

However, although these described challenges were not systematically addressed and resolved within the methodology, spatial interpretation and manual examination of selected outliers provided valuable insights into the context in which these anomalies occurred. This topic in itself is intriguing, even though it somewhat deviates from the original research question concerning vernacular fuzzy places and geographies.

6.1.3 Sales language of listing descriptions

A significant challenge encountered was the "sales language" used in the listings. Especially prevalent terms like "central location" and "quiet location," as well as city-wide known nodes such as Paradeplatz or the main train station, and natural landmarks like Uetliberg or Lake Zurich, were employed as selling points so extensively across the entire city area that their spatial patterns appeared almost devoid of value.

In this context, it is important to note that the extraction of toponyms from listing texts for the mapping of vernacular places and areas might be better suited for less frequently mentioned place names and location references. These are less likely to be exploited as "selling points" and maintain a tighter geographical relevance.

While this implies that the research questions about the footprints of perceived geographies are not appropriate for frequently cited location references, it is still possible to make intriguing observations regarding the importance of a place. This is evident with the airport, which was excluded from the analysis due to its location outside the city, yet still ranked among the most frequently mentioned places. The more significant a place/landmark/node is, the larger the area in which it is mentioned. Lake Zurich or the main train station with its many connections serves as a selling point for the city as a whole and is therefore cited throughout the city. In contrast, e.g., Goldbrunnenplatz, being smaller, serves as an essential transport hub in the district and hence is mentioned in a more localized context.

This high occurrence of “selling language” inevitably raises the question of whether the classification of real estate data as VGI (Volunteered Geographic Information) based on existing literature (Hu et al., 2019) might warrant re-evaluation. After all, this data is not strictly "volunteered" in terms of geographic information. The listing of a housing unit is made with the intention of placing it on the open market, and as such, it isn't 100% voluntary. If someone were to describe the location of a dwelling purely voluntarily, the description would likely be less dominated by embellishments aimed at attracting prospective tenants, even if they aren't entirely accurate.

6.1.4 Limitation of spatial analysis through KDE

The results of a KDE (Kernel Density Estimation) are influenced less by the exact methodology than by the used kernel size or bandwidth size. The present study utilized a kernel size of 100m, as this made sense especially in smaller areas for better visibility. However, using other adaptive approaches as demonstrated by Hollenstein and Purves (Hollenstein & Purves, 2010) might have yielded more meaningful results in certain cases.

6.1.5 Research questions

Given the aforementioned limitations, the initial research questions can be addressed as follows:

1. *Which location references are of utmost relevance in Zurich-City?*

The most prevalent location reference in Zurich residential property advertisements is, unsurprisingly, the name of the city itself: "Zurich" (aggregated from various spellings such as "Zürich", "Züri" (Swiss German), "Zuerich", "Stadt", and "Stadt Zürich/Zuerich/etc.") that occurred 34,606 times.

However, Zurich, along with other frequently occurring toponyms outside of the city such as major large cities (Basel, Bern, etc.) and the airport in the municipality of Kloten, was excluded. Consequently, the most common location references for each examined category are as follows:

District

The most frequently mentioned districts were, for the most part, also officially recognized districts. An exception, for instance, was the fuzzy region "Innenstadt" (Inner City). Following are the top 20 districts listed in descending order based on their number of mentions for all districts:

Oerlikon (6,181), Seebach (4,256), Altstetten (3,933), Höngg (2,854), Wollishofen (2,798), Seefeld (2,425), Witikon (2,294), Wiedikon (2,252), Affoltern (1,825), Kreis 6 (1,789), Kreis 4 (1,593), Wipkingen (1,514), Altstadt (1,486), Innenstadt (1,469), Albisrieden (1,441), Kreis 3 (1,332), Schwamendingen (1,290), Kreis 7 (1,037), Niederdorf (996), Leimbach (921).

Nodes (incl. Nodes of Public Transport)

For the nodes, which are combined here with the special case of Public transport, the major train stations were predominantly mentioned. The dominance of train stations among the nodes underscores the importance of accessibility to a railway station, even in the presence of other public transportation options. This is also reflected in the fact that access to public transport has a value-enhancing effect on real estate (Cordera et al., 2019).

Thus, the top four most mentioned nodes are all train stations. Following are the top 20 nodes listed in descending order based on their number of mentions:

Main station (9,514), Oerlikon station (6,206), Altstetten station (2,286), Stadelhofen station (1,916), Paradeplatz (1,491), Wiedikon station (1,432), Hardbrücke station (1,390), Enge station (1,321), Kreuzplatz (1,178), Limmatplatz (1,160), Goldbrunnenplatz (975), Farbhof (778), Meierhofplatz (765), Römerhof (760), Schaffhauserplatz (756), Wipkingen station (735), Berninaplatz (728), Tiefenbrunnen station (695), Schwamendingerplatz (690), and Wollishofen station (685).

Landmarks (including natural landmarks)

Among the significant landmarks, it is not necessarily tall buildings that serve for orientation in the sense of Lynch (1960), such as the Prime Tower, Lochergut, Hardau Towers, or other skyscrapers (see also section 2.4.2). Instead, it the majority is about landmarks with a "utility", such as recreational areas (Uetliberg, Katzensee), shopping centers (Sihlcity, Letzipark), other places for leisure activities (indoor swimming pools), or views (Alps, Lake Zurich).

Following are the top 20 landmarks listed in descending order based on their number of mentions: Lake of Zurich (6,318), Uetliberg (5,652), Zürichberg (2,787), Limmat (2,680), Letzipark (1,732), Sihlcity (1,634), Triemli (1,350), indoor swimming pool Oerlikon (1,320), Alps (1,271), Irchelpark (1,184), Hönggerberg (1,016), University (932), University of Irchel (907), Katzensee (775), Käferberg (757), Swimming pool Allenmoos (726), Glatt (698), university hospital (622), Werdinsel (485), and Sihl (474).

Descriptions

The descriptions were primarily derived from the n-gram extracted descriptions associated with the word "location." They only formed a small group within the considered 125 locations with 200 or more entries, but still "central location and "quiet location" were the two locations with most entries. The following are listed in descending order: central location ("zentrale Lage", 25,943), quiet location ("ruhige Lage", 14,597), greenery ("Grünen", 651, can appear as phrases like "quickly in the greenery"), prime residential location ("beste Wohnlage", 504), proximity to the center ("Zentrumsnähe", 300), and on the edge of the forest ("am Waldrand", 285).

2. *What spatial distribution patterns are typical for these location references, how do they possibly differentiate from official boundaries and locations, and how can they be explained?*

The generated footprints vary depending on the number of occurrence of terms. Generally, it can be asserted that toponyms or place references that occur less frequently tend to provide a more geographically accurate footprint. The more frequently they appear, the broader the geographic area in which they are used, thus giving more insight into the importance perimeter rather than vernacular geography of the given location.

Districts

Regarding districts, there are varying examples of footprints. For officially recognized districts, which can be juxtaposed with official district boundaries, the predominant trend is that their extracted designations occur within the geographical confines of the district. However, district boundaries are often not precisely adhered to, especially where they are fluid and do not follow a "clear demarcation". An issue here is that prominent and popular districts are used extensively beyond their boundaries, as seen with the prestigious Seefeld, which comprehensively incorporates the neighboring district of Mühlebach. Naturally, there are sporadic kernels at large distances which seem illogical. The reasons for these are manifold. Either a connection to the concerned district is described or a view of it (e.g., the frequent mention of Altstetten in Höngg due to Höngg's elevated location providing a vista over Altstetten). The unofficial district of Innenstadt or Altstadt is extensively used throughout the city, as it serves as a "selling point" (e.g., "quickly accessible Zurich city center or historic old town").

Nodes and nodes of public transport

The spatial footprints for nodes also exhibit significant variability depending on their prominence. The most frequently mentioned Paradeplatz is referenced throughout the city. In contrast, smaller, locally significant nodes, such as the Goldbrunnenplatz, are most often mentioned within and immediately around the vicinity of the node itself. Here too, there are isolated kernels at a greater distance, which often describe a transport link. For squares that serve as crucial transportation hubs, the density map often highlights the areas where they hold importance. For instance, with Bucheggplatz, where multiple tram and bus lines operate at high frequencies, the density map predominantly extends in the direction of Affoltern. This area often finds its connection via the Bucheggplatz since it doesn't have its train station. However, in the neighboring district of Oerlikon, mentions drop off abruptly. This is because Oerlikon has better transportation infrastructure, particularly with its Oerlikon train station, which is referenced much more frequently than Bucheggplatz.

Landmarks and natural Landmarks

Regarding landmarks, a universal spatial pattern cannot be deduced; rather, the pattern significantly depends on their prominence and relevance. For instance, "nearby" amenities such as the Sihlcity shopping center or parks like Rieterpark are mainly mentioned within their immediate vicinity, where they can be highlighted as neighborhood assets. Particularly, natural landmarks are referenced more expansively. The Lake Zurich, for example, is ubiquitously advertised around the entire lake basin and even at considerably farther distances, spanning across the entire city or in elevated locations where views of the lake can be promoted. The same pattern is observed with Uetliberg. The Alps, being a frequently mentioned natural landmark, distinctly pertain to the panoramic views they offer and hence are predominantly featured in advertisements for properties located on sloped terrains with mountain vistas.

Descriptions

The descriptive terms were only useful when they weren't overused. Classic descriptors such as "central location" and "quiet location" appear so frequently that they are almost universally applied throughout the city, clearly establishing themselves as selling points. In contrast, more specific descriptions like " on the edge of the forest" were less prevalent. Although this particular description did not generate a large footprint, the kernels were indeed closely associated with forested areas among Uetli- and Zürichberg.

6.2 Implications

The present study distinguishes itself from previous research in several ways, thereby offering a complement to the existing body of work. The distinguishing characteristics, compared to prior studies, arise particularly from the combination of the following factors:

- language NER was provided on (German)
- geographical location studied (Zurich, Switzerland)
- used approach / methodology (systematic examination of the entire dataset for all possible spatial references)
- the data the analysis was conducted on itself, as it was a) a dataset of real estate listings and b) spanning an extended period of nearly 20 years
- and last, it was conducted in a challenging residential market environment characterized by a housing shortage

While preceding studies related to real estate predominantly emphasized the advantages of such data, this thesis has particularly brought to light the associated challenges and problems. Other insights in this context are only known from a very recent study by Brunila et al. (2023), although it did not focus on traditional residential properties but on AirBnbs, which might be closer to hotel sector than to residential real estate. In that research, reviews were compared with AirBnb listings, revealing that the term "near" was exaggeratedly used to make the AirBnb seem more appealing. It appears that a similar trend exists for residential listings, even in a market facing a housing shortage.

Thus, this study has highlighted the challenges associated with considering real estate listings as VGI. Given that these listings do not originate from voluntary users but serve a specific purpose (renting/selling), the language used often presents the property in the best possible light, which leads to great exaggerations, that impact the spatial distribution of a used term.

The geographical footprints, when viewed in the context of vernacular geography, were thus rendered meaningless in many cases. However, the extracted location references did still provide insights into what location references are perceived as valuable in Zurich's real estate market and footprints were useful for less-used toponyms on a smaller geographical level.

It is also worth noting that the timing of this study was fortuitous. The recent emergence of text-generating AI tools for the general public, such as ChatGPT, suggests that listing texts may soon not be of human origin. However, as human-generated content is essential for tasks involving natural language processing (Liddy, 2001), this thesis could not have been conducted later, or only when a exclusion of newly-added AI-texts would be made.

7 Conclusion

7.1 Main findings

In conclusion, the following insights can be summarized:

1. The data foundation of real estate listings is influenced in its scope by the tense market environment. This results in many apartments being passed on informally and therefore not being listed, thus not appearing in the dataset.

2. The extent to which this tense market environment has had a reducing influence on the remaining descriptions cannot be conclusively answered with the methods used.

3. The extracted location references with the most hits relate to broadly applied generic descriptions (central & quiet location), to significant transit hubs associated with public transport, natural landmarks (such as mountains and bodies of water), large city districts, and well-known squares.

4. The creation of spatial footprints for the extraction methods used via KDE is not equally suitable for all toponyms/location references when considering the quantification of vernacular geographies. This is especially the case when they are excessively used as selling points, or if they need to be viewed in a more differentiated manner in terms of the linguistic context used (e.g., whether the location itself is being described or a connection to a place).

5. Nonetheless, insights can be derived from some of the KDE's, such as the perimeter within which a transit hub is essential and the point at which another becomes more important, from which locations one has a view of another location, or for smaller nodes, the radius within which something is perceived as belonging to a place.

7.2 Future Work

Despite the initially described weaknesses, the dataset provides an exciting foundation for further work in the area of Geographic Information Retrieval (GIR).

To truly get to the heart of vernacular geographies, toponyms could be extracted alongside their mentioned context. This would allow differentiation between the actual location and Points of Interest (POI) in the vicinity. Thus, further intriguing questions about vernacular geography (location) and spatial related expressions (like "near") could be posed.

It would also be interesting to incorporate a temporal aspect. For instance, research could examine how the attractiveness of certain districts has evolved over time. This could be gauged either through the number of mentions or through textual analysis methods like cosine similarity, examining which locations are described as trendy and urban, and from what point in time. However, a challenge here, of course, would be that the texts might portray a location more positively than its actual reputation. Even an area known publicly as a less desirable place at a certain point in time would not necessarily be reflected as such in the listing text.

8 Bibliography

Acheson, E., & Purves, R. S. (2021). Extracting and modeling geographic information from scientific articles. *PLoS ONE, 16*(1 January). https://doi.org/10.1371/journal.pone.0244918

Aeberli, M. (2023, August 14). În Zürich eine neue Wohnung zu finden, ist ein Vollzeitjob. *Tagesanzeiger*. https://www.tagesanzeiger.ch/in-zuerich-eine-neue-wohnung-zu-finden-ist-ein-vollzeitjob-494396952828

Akoglu, H. (2018). User's guide to correlation coefficients. In *Turkish Journal of Emergency Medicine* (Vol. 18, Issue 3, pp. 91–93). Emergency Medicine Association of Turkey. https://doi.org/10.1016/j.tjem.2018.08.001

Amt für Städtebau, & Stadt Zürich. (2022). *Stadt Zürich Hochhäuser*. https://hochhaeuser.stadt-zuerich.ch/

Ardanuy, M. C., & Sporleder, C. (2017). Toponym disambiguation in historical documents using semantic and geographic features. *ACM International Conference Proceeding Series, Part F129473*, 175–180. https://doi.org/10.1145/3078081.3078099

Bahrehdar, A. R., Adams, B., & Purves, R. S. (2020). Streets of London: Using Flickr and OpenStreetMap to build an interactive image of the city. *Computers, Environment and Urban Systems, 84*. https://doi.org/10.1016/j.compenvurbsys.2020.101524

Batista, D. S., Ferreira, J. D., Couto, F. M., & Silva, M. J. (2012). Toponym Disambiguation Using Ontology-Based Semantic Similarity. In H. Caseli, A. Villavicencio, A. Teixeira, & F. Perdigao (Eds.), *Computational Processing of the Portguese Language. PROPOR 2012.* (PROPOR 2012, Vol. 7243). Springer.

Baumann, A. (2019). Multilingual Language Models for Named Entity Recognition in German and English. *Proceedings of the Student Research Workshop Associated with RANLP 2019*, 21–27. https://doi.org/10.26615/issn.2603-2821.2019_004

Brunila, M., Verma, P., & Mckenzie, G. (2023). When Everything Is ``Nearby'': How Airbnb Listings in New York City Exaggerate Proximity. *12th International Conference on Geographic Information Science (GIScience 2023)*. https://doi.org/10.4230/LIPIcs.GIScience.2023.16

Brunner, T. J., & Purves, R. (2008). Spatial Autocorrelation and Toponym Ambiguity. *GIR '08 : Proceedings of the 5th Workshop on Geographic Information Retrieval*.

Buscaldi, D. (2011). Approaches to disambiguating toponyms. *SIGSPATIAL Special*, *3*(2), 16–19. https://doi.org/10.1145/2047296.2047300

Cambridge Dictionary. (2023). information retrieval. In *Cambridge Dictionary*. Cambridge University Press & Assessment.

Chen, X., & Biljecki, F. (2022). Mining real estate ads and property transactions for building and amenity data acquisition. *Urban Informatics*, *1*(1), 12. https://doi.org/10.1007/s44212-022-00012-2

Chesnokova, O., & Purves, R. S. (2018, November 6). Automatically creating a spatially referenced corpus of landscape perception. *Proceedings of the 12th Workshop on Geographic Information Retrieval, GIR 2018*. https://doi.org/10.1145/3281354.3281356

Chiche, A., & Yitagesu, B. (2022). Part of speech tagging: a systematic review of deep learning and machine learning approaches. In *Journal of Big Data* (Vol. 9, Issue 1). Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1186/s40537-022-00561-y

Chieu, H. L., & Ng, H. T. (2002). Named Entity Recognition: A Maximum Entropy Approach Using Global Information. *The 19th International Conference on Computational Linguistics*. http://maxent.sourceforge.net

Chinchor, N. (1997). MUC-7 Named Entity Task Definition. *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Viriginia, April 29 - May 1, 1998*.

Coleman, D. J., Georgiadou, Y., & Labonte, J. (2009). Volunteered Geographic Information: The Nature and Motivation of Produsers *. *International Journal of Spatial Data Infrastructures Research*, *4*, 332–358. https://doi.org/10.2902/1725-0463.2009.04.art16

Collins English Dictionary. (2023). information retrieval. In *Collins English Dictionary*. HarperCollins Publishers. https://www.collinsdictionary.com/de/worterbuch/englisch/information-retrieval#google_vignette

Cordera, R., Coppola, P., dell'Olio, L., & Ibeas, Á. (2019). The impact of accessibility by public transport on real estate values: A comparison between the cities of Rome and Santander. *Transportation Research Part A: Policy and Practice*, *125*, 308–319. https://doi.org/10.1016/j.tra.2018.07.015

Cresswell, Tim. (2004). *Place : a short introduction*. Blackwell Pub.

Davies, C., Holt, I., Green, J., Harding, J., & Diamond, L. (2009). User needs and implications for modelling vague named places. *Spatial Cognition and Computation*, *9*(3), 174–194. https://doi.org/10.1080/13875860903121830

Derungs, C., & Purves, R. S. (2014). From text to landscape: locating, identifying and mapping the usse of landscape features in a Swiss Alpine Corpus. *International Journal of Geographical Information Science*, *28*(6), 1272–1293.

Derungs, C., & Purves, R. S. (2016). Mining nearness relations from an n-grams Web corpus in geographical space. *Spatial Cognition and Computation*, *16*(4), 301–322. https://doi.org/10.1080/13875868.2016.1246553

Duckham, M., & Worboys, M. (2001). *Computational structure in three-valued nearness relations*.

Egenhofer, M. J., & Mark, D. M. (1995). *Naive Geography **.

Eichler, M., Jank, K., Kraml, A., Röser, P., & Rufer, R. (2013). *Raumkonzept und Handlungsräume in der Schweiz in Zahlen*. www.bakbasel.com

Ettlin, L. (2011). *Identifying Unregistered Vernacular Toponyms and Alternative Spellings of Placenames in Switzerland*. University of Zurich.

Evans, A. J., & Waters, T. (2007). Mapping vernacular geography: web-based GIS tools for capturing "fuzzy" or "vague" entities. In *Int. J. Technology, Policy and Management* (Vol. 7, Issue 2).

Federal Statistical Office. (2022). Leer stehende Wohnungen nach Grossregion, Kanton, Gemeinde, Anzahl Wohnräumen und Typ der Leerwohnung. In *Leerwohnungszählung*. https://www.bfs.admin.ch/bfs/de/home/statistiken/bau-wohnungswesen/wohnungen/leerwohnungen.assetdetail.23404272.html

Federal Statistical Office, STATPOP, & eidgenössiche Volkszählung. (2022). *Personen in Privathaushalten nach Haushaltsgrösse, 1930-2021*. https://www.bfs.admin.ch/bfs/de/home/statistiken/bevoelkerung/stand-entwicklung/haushalte.html

Fogliaroni, P., & Hobel, H. (2015). *Implementing Naive Geography via Qualitative Spatial Relation Queries*.

Frey, U., & Stadt Zürich. (2022). *Mietpreise in der Stadt Zürich*. https://www.stadt-zuerich.ch/prd/de/index/statistik/publikationen-angebote/publikationen/webartikel/2022-11-03_Mietpreise-in-der-Stadt-Zuerich.html

Fuchs, M. (2008). Langstrasse: Quartier im Umbruch. In *Studie zur Entwicklung im Langstrassenquartier*. University of Zurich. https://www.news.uzh.ch/de/articles/2008/3212.html

Gelernter, J., & Mushegian, N. (2011). Geo-parsing messages from microtext. *Transactions in GIS*, *15*(6), 753–773. https://doi.org/10.1111/j.1467-9671.2011.01294.x

Goodchild, M. (2007). Citizens as sensors: The world of volunteered geography. In *GeoJournal* (Vol. 69, Issue 4, pp. 211–221). https://doi.org/10.1007/s10708-007-9111-y

Goodchild, M. F., & Hill, L. L. (2008). Introduction to digital gazetteer research. In *International Journal of Geographical Information Science* (Vol. 22, Issue 10, pp. 1039–1044). https://doi.org/10.1080/13658810701850497

Goodchild, M., & Li, L. (2012). Assuring the quality of volunteered geographic information. *Spatial Statistics*, *1*, 110–120. https://doi.org/10.1016/j.spasta.2012.03.002

Goyal, A., Gupta, V., & Kumar, M. (2018). Recent Named Entity Recognition and Classification techniques: A systematic review. In *Computer Science Review* (Vol. 29, pp. 21–43). Elsevier Ireland Ltd. https://doi.org/10.1016/j.cosrev.2018.06.001

Granell, C., & Ostermann, F. O. (2016). Beyond data collection: Objectives and methods of research using VGI and geo-social media for disaster management. *Computers, Environment and Urban Systems*, *59*, 231–243. https://doi.org/10.1016/j.compenvurbsys.2016.01.006

Grishman, R., & Sundheim, B. (1996). Message Understanding Conference-6: A Brief History. *Proceedings of the 16th International Conference on Computational Linguistic*, *I*, 466–471.

Grothe, C., & Schaab, J. (2009). Automated footprint generation from geotags with kernel density estimation and support vector machines. *Spatial Cognition and Computation*, *9*(3), 195–211. https://doi.org/10.1080/13875860903118307

Grover, C., Givon, S., Tobin, R., & Ball, J. (2008). Named Entity Recognition for Digitised Historical Texts. *Proceedings of Sixth International Conference on Language Resources and Evaluation (LREC'08)*. www.ltg.ed.ac.uk/software/ltxml2

Hall, M. M., & Jones, C. B. (2022). Generating geographical location descriptions with spatial templates: a salient toponym driven approach. *International Journal of Geographical Information Science*, *36*(1), 55–85. https://doi.org/10.1080/13658816.2021.1913498

Hall, M. M., Smart, P. D., & Jones, C. B. (2011). *Interpreting Spatial Language in Image Captions*. http://www.flickr.com

Hill, L. L. (2006). Gazetteers and Gazeteer Services. In *Georeferencing: The Geographic Assocations of Information*. Cambridge, Mass. : MIT Press.

Hollenstein, L. (2008). *Capturing Vernacular Geography from Georeferenced Tags*.

Hollenstein, L., & Purves, R. S. (2010). Exploring place through user-generated content: Using Flickr tags to describe city cores. *Journal of Spatial Information Science*, *1*(2010), 21–48. https://doi.org/10.5311/JOSIS.2010.1.3

Hu, Y. (2018). Geo-text data and data-driven geospatial semantics. *Geography Compass*, *12*(11). https://doi.org/10.1111/gec3.12404

Hu, Y., Mao, H., & McKenzie, G. (2019). A natural language processing and geospatial clustering framework for harvesting local place names from geotagged housing advertisements. *International Journal of Geographical Information Science*, *33*(4), 714–738. https://doi.org/10.1080/13658816.2018.1458986

Ishikawa, T., & Montello, D. R. (2006). Spatial knowledge acquisition from direct experience in the environment: Individual differences in the development of metric knowledge and the integration of separately learned places. *Cognitive Psychology*, *52*(2), 93–129. https://doi.org/10.1016/j.cogpsych.2005.08.003

Jones, C. B., Alani, H., & Tudhope, D. (2001). Geographical Information Retrieval with Ontologies of Place. In D. R. Montello (Ed.), *Spatial Information Theory* (Vol. 2205).

Jones, C. B., & Purves, R. S. (2008). Geographical information retrieval. In *International Journal of Geographical Information Science* (Vol. 22, Issue 3, pp. 219–228). https://doi.org/10.1080/13658810701626343

Jones, C. B., Purves, R. S., Clough, P. D., & Joho, H. (2008). Modelling vague places with knowledge from the Web. *International Journal of Geographical Information Science*, *22*(10), 1045–1065. https://doi.org/10.1080/13658810701850547

Kanton Zürich. (2023). *Gemeindeporträt*. Gemeindeporträt. https://www.zh.ch/de/politik-staat/gemeinden/gemeindeportraet.html

Kessler, C., Jonowicz, K., & Bishr, M. (2009). An agenda for the next generation gazetteer: geograhic information contribution and retrieval. *GIS '09: Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 91–100.

Kherwa, P., & Bansal, P. (2020). Topic Modeling: A Comprehensive Review. *EAI Endorsed Transactions on Scalable Information Systems*, *7*(24), 1–16. https://doi.org/10.4108/eai.13-7-2018.159623

Klimek, B., Ackermann, M., Kirschenbaum, A., & Hellmann, S. (2017). Investigation the Morphological Complexity of German Named Entities: The Case of GermEval NER Challenge. In G. Rehm & T. Declerck (Eds.), *Language Technologies for the Challenges of the Digital Age*. Springer. http://www.springer.com/series/1244

Liddy, E. D. (2001). Natural Language Processing. In *Encyclopedia of Library and Information Science: Vol. 2nd Ed*. https://surface.syr.edu/istpub

Lieberman, M. D., & Hanan, S. (2011). Multifaceted Toponym Recognition for Streaming News. *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1346.

Lieberman, M. D., Samet, H., & Sankaranaarayanan, J. (2010). Geotagging with Local Lexicons to Build Indexes for Textually-Specified Spatial Data. *Data Engineering (ICDE), 2010 IEEE 26th International Conference*.

Lynch, K. (1960). *The Image of the City*.

Manning, C. D., Raghavan, P., & Schütze, H. (2009). *An Introduction to Information Retrieval* (Online edition (c)). Cambridge University Press.

Meier, C. (2023, April 10). Zürcher Mieter machen Verwaltungen unmoralische Angebote. *Nau.Ch*. https://www.nau.ch/news/schweiz/zurcher-mieter-machen-verwaltungen-unmoralische-angebote-66465425

Ménard, D. (2023, August 3). *Steckt die Schweiz in einer selbst gemachten Wohnungsnot?* thebranch.

Metzler, B. (2023, May 26). Wer bekommt die 300 Millionen Franken aus dem Wohnraumfonds? *Tagesanzeiger*. https://www.tagesanzeiger.ch/wer-bekommt-die-300-millionen-franken-aus-dem-wohnraumfonds-777369086870

Montello, D. R. (1998). *A New Framework for Understanding the Acquisition of Spatial Knowledge in Large-Scale Environments* (R. G. Goodchild & M. J. Egenhofer, Eds.). Oxford University Press.

Montello, D. R. (2009). Navigation. In *The Cambridge Handbook of Visuospatial Thinking* (pp. 257–294). Cambridge University Press. https://doi.org/10.1017/cbo9780511610448.008

Montello, D. R., Goodchild, M. F., Gottsegen, J., & Fohl, P. (2003). Where's downtown?: Behavioral methods for determining referents of vague spatial queries. *Spatial Cognition and Computation*, *3*(2–3), 185–204. https://doi.org/10.1080/13875868.2003.9683761

Mooers, C. N. (1950). The Theory of Digital Handling of Non-Numerical Information and Its Implications to Machine Economomics. *Technical Bulletin*, *48*.

Mooney, P., & Minghini, M. (2017). A Review of OpenStreetMap Data. In G. Foody, L. See, S. Fritz, P. Mooney, A.-M. Olteanu-Raimon, C. Fonte Costa, & V. Antoniou (Eds.), *Mapping and the citizen senor* (pp. 37–59). Ubiquity Press.

Nilsson, I., & Delmelle, E. C. (2023). An embedding-based text classification approach for understanding micro-geographic housing dynamics. *International Journal of Geographical Information Science*. https://doi.org/10.1080/13658816.2023.2209803

Pasley, R., Clough, P., Purves, R. S., & Twaroch, F. A. (2008). Mapping geographic coverage of the web. *GIS: Proceedings of the ACM International Symposium on Advances in Geographic Information Systems*, 154–162. https://doi.org/10.1145/1463434.1463459

Purves, R. S., Clough, P., Jones, C. B., Hall, M. H., & Murdock, V. (2018). Geographic information retrieval: Progress and challenges in spatial search of text. In *Foundations and Trends in Information Retrieval* (Vol. 12, Issues 2–3, pp. 164–318). Now Publishers Inc. https://doi.org/10.1561/1500000034

Rauch, E., Bukatin, M., & Baker, K. (2003). *A confidence-based framework for disambiguating geographic terms*. 50–54. https://doi.org/10.3115/1119394.1119402

Realmatch360. (2023). *Municipal report Zurich-City*.

Rey, U. (2020, September 29). *Hohe Wohnungsfluktuation trotz tiefem Leerstand*. Stadt Zürich Präsidialdepartement. https://www.stadt-zuerich.ch/prd/de/index/statistik/publikationen-angebote/publikationen/webartikel/2020-09-29_Hohe-Wohnungsfluktuation-trotz-tiefem-Leerstand.html#

Sanderson, M., & Croft, W. B. (2012). The history of information retrieval research. *Proceedings of the IEEE*, *100*(SPL CONTENT), 1444–1451. https://doi.org/10.1109/JPROC.2012.2189916

Schmitt, X., Kubler, S., Robert, J., Papadakis, M., & LeTraon, Y. (2019). A Replicable Comparison Study of NER Software: StanfordNLP, NLTK, OpenNLP, SpaCy, Gate. *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, 338–343.

Schockaert, S., De Cock, M., Cornelis, C., & Kerre, E. E. (2008). Fuzzy region connection calculus: Representing vague topological information. *International Journal of Approximate Reasoning*, *48*(1), 314–331. https://doi.org/10.1016/j.ijar.2007.10.001

spaCy. (2023). *spaCy 101: Everything you need to know*. https://spacy.io/usage/spacy-101

Stadt Zürich Statistik. (2023a). *Bevölkerungsdichte nach Stadtkreis und Stadtquartier*. https://www.stadt-zuerich.ch/prd/de/index/statistik/themen/bevoelkerung/bevoelkerungsentwicklung/bevoelkerungsdichte.html

Stadt Zürich Statistik. (2023b). *Bevölkerungsentwicklung nach Stadtkreis und Stadtquartier seit 1930*. https://www.stadt-zuerich.ch/prd/de/index/statistik/themen/bevoelkerung/bevoelkerungsentwicklung/kreise-und-quartiere.html

Stadt Zürich Statistik. (2023c). *Quartierspiegel Escher Wyss*. https://www.stadt-zuerich.ch/prd/de/index/statistik/publikationen-angebote/publikationen/Quartierspiegel/QUARTIER_052.html

Stadt Zürich Statistik. (2023d). *Umzüge innerhalb der Stadt nach Herkunft, 1971-2022*. https://www.stadt-zuerich.ch/prd/de/index/statistik/themen/bevoelkerung/zuzug-wegzug-umzug/umzuege.html#daten

Stadt Zürich Statistik. (2023e). *Zürcher Index der Konsumentenpreise, Mietpreisindex, Monatswerte seit August 1939*. https://www.stadt-zuerich.ch/prd/de/index/statistik/themen/bauen-wohnen/mietpreise/mietpreisindexerhebung.html

Stadt Zürich Statistik. (2023f). *Zuzüge von auswärts nach Herkunft und Geschlecht, seit 1971*. https://www.stadt-zuerich.ch/prd/de/index/statistik/themen/bevoelkerung/zuzug-wegzug-umzug/zuzuege.html#daten

Stadt Zürich Statistik, & GWZ. (2023). *Wohnungsbestand nach Zimmerzahl seit 1930*. https://www.stadt-zuerich.ch/prd/de/index/statistik/themen/bauen-wohnen/gebaeude-wohnungen/wohnungsbestand.html

Statistik Stadt Zürich. (2005). *Zürich - eine Stadt wie jede andere? Schweizer Städte und Agglomerationen zwischen 1970 und 2000*. www.statistik-stadt-zuerich.info

Strobel, J. (2013, October 9). Irgendwann kam die Verzweiflung. *Tagesanzeiger*. https://www.tagesanzeiger.ch/irgendwann-kam-die-verzweiflung-469161662283

Swanson, D. R. (1998). Historical Note: Information Retrieval and the Future of an Illusion. *Journal of the American Society for Information Science 39(2).*

Tagesanzeiger. (2019). Hunderte stehen Schlange für 1148-Franken-Wohnung. *Tagesanzeiger*. https://www.tagesanzeiger.ch/hunderte-stehen-schlange-fuer-1148-franken-wohnung-231250633014

Tata, S., & Patel, J. M. (2007). Estimating the Selectivity of tf-idf based Cosine Similarity Predicates. *ACM SIGMOD Record*, *36*(2), 7–12.

Teitler, B. E., Lieberman, M. D., Panozzo, D., Sankaranarayanan, J., Samet, H., & Sperling, J. (2008). NewsStand: A New View on News*. *GIS '08 Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 1–10.

Thalmann P. (2012). Housing Market Equilibrium (almost) without Vacancies. *Urban Studies, 49*(8), 1643–1658. https://doi.org/10.2307/26150951

Tobler, W. R. (1970). A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography*, *46*, 234–240.

Twaroch, F. A., Jones, C. B., & Abdelmoty, A. I. (2008). Acquisition of a vernacular gazetteer from web sources. *LOCWEB '08: Proceedings of the First International Workshop on Location and the Web*, 61–64.

Twaroch, F. A., Purves, R. S., & Jones, C. B. (2009). *Stability of Qualitative Spatial Relations between Vernacular Regions Mined from Web Data*. www.gumtree.com

Vakaridis, M. (2023). Wie Google die Mieten in Zürich zum Explodieren bringt. *SWI Swissinfo.Ch*. https://www.swissinfo.ch/ger/wirtschaft/wie-google-die-mieten-in-zuerich-zum-explodieren-bringt/48555244

Von Ledebur, M. (2023, March 25). Reisst die Credit Suisse das geplante neue Zürcher Fussballstadion auf dem Hardturm mit in den Abgrund? *Neue Zürcher Zeitung*. https://www.nzz.ch/zuerich/reisst-die-credit-suisse-das-geplante-neue-zuercher-fussballstadion-auf-dem-hardturm-mit-in-den-abgrund-ld.1732064?reduced=true

Wallgrün, J. O., Klippel, A., & Baldwin, T. (2014, November 4). Building a corpus of spatial relational expressions extracted from web documents. *Proceedings of the 8th Workshop on Geographic Information Retrieval, GIR 2014*. https://doi.org/10.1145/2675354.2675702

Webster, J. J., & Kit, C. (1992). TOKENIZATION AS THE INITIAL PHASE IN NLP. *The 14th International Conference on Computational Linguistics*, *4*.

Zahra, K., Ostermann, F., Deb Das, R., & Purves, R. (2022, May). Towards an Automated Information Extraction Model from Twitter Threads during Disasters. *WiP Paper - Social Media for Crisis Management*. https://www.researchgate.net/publication/365769575

Zelianskaia, N. L., Belousov, K. I., Galinskaia, T. N., & Ichkineeva, D. A. (2020). Naive geography: geoconceptology and topology of geomental maps. *Heliyon*, *6*(12). https://doi.org/10.1016/j.heliyon.2020.e05644

Zürich Tourismus. (2023). *Die 12 Stadtteile von Zürich*. https://www.zuerich.com/de/besuchen/ueber-zuerich/zuerichs-stadtteile

Personal declaration

I hereby declare that the submitted thesis is the result of my own, independent work. All external sources are explicitly acknowledged in the thesis.

Zurich, 30th September 2023

Chantal Meier