universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„Möglichkeiten zur Lokalisation von Standorten anhand von quantitativen Entfernungsangaben in Texten mittels Textmining am Beispiel Airbnb“

verfasst von / submitted by

Lukas Wilfinger, MSc BSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2019 / Vienna 2019

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 066 856
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Kartographie und Geoinformation
Betreut von / Supervisor::	Ass.-Prof. Mag. Dr. Andreas Riedl

Inhalt

Inhalt ..... **i**

Abbildungsverzeichnis ..... **iv**

Tabellenverzeichnis ..... **v**

Abkürzungsverzeichnis ..... **vi**

Kurzfassung ..... **vii**

Abstract ..... **viii**

1 Einleitung **1**

1.1 Problemstellung 1

1.2 Zielsetzung 2

1.3 Relevanz der Arbeit 2

1.4 Forschungsfragen 3

1.5 Verwendete Methodik und Aufbau der Arbeit 4

2 Stand der Forschung **6**

2.1 Data Mining 6

2.1.1 Prinzipien 6

2.1.2 Vorgehensweise 9

2.1.3 Textmining 10

2.1.4 Räumliches Data Mining 11

2.2 Natürliche Sprachprozessierung 13

2.3 Kombination von GIS und Text 16

2.4 Räumliche Information aus textuellen Daten 19

2.4.1 Geosemantik und Geoparsing 19

2.4.2 Fuzzy-Logik beim Geoparsing 24

2.4.3 Schwierigkeiten mit nutzergenerierten Daten 25

2.4.4 Räumliche Beziehungen in der Sprache 27

2.5 Menschliche Wahrnehmung von Distanzen 29

2.6 Erreichbarkeitsanalysen 33
2.7 Zwischenfazit 35
3 Datenquellen 37
3.1 Airbnb 37
3.2 Inside Airbnb 38
3.2.1 Rechtliche Grundlage 39
3.2.2 Aufbereitung der Daten 39
3.2.3 Qualität von nutzergenerierten Informationen 40
3.2.4 Festlegung der Anforderungen 42
3.3 Graphenintegrations-Plattform 44
3.4 Open Data der Stadt Wien 44
3.5 Frei verfügbare Informationen und Technologien 45
4 Methodik zur Standortermittlung 46
4.1 Vorgehensweise 46
4.2 Untersuchungsraum 48
4.2.1 Auswahl einer Stichprobe 49
4.2.2 Datenkollektion für die Untersuchung 50
4.3 Verwendete Technologien 51
4.3.1 R. 51
4.3.2 GIS-Software 53
4.4 Praktische Aufarbeitung 54
4.4.1 Wichtige Paradigmen in R 54
4.4.2 Extraktion räumlich relevanter Information 55
4.4.3 Ethischer Grundgedanke 63
5 Implementierung in ein GIS 64
5.1 Erforderliche Vorarbeiten 64
5.2 Auswahl der treffendsten Distanzangabe 65
5.3 Ergebnisse 65

5.3.1Erstellung des Einzugsgebiets 66

5.4 Messen und Vergleichen der Ergebnisse 68

5.5 Bewertung 71

6 Diskussion 75

7 Zusammenfassung und Ausblick 80

8 Literatur 83

Anhang I R-Code zur Extraktion von räumlichen und zeitlichen Distanzangaben aus den Bewertungen 93

Anhang II R-Code zur Extraktion von räumlichen und zeitlichen Distanzangaben aus den Beschreibungen 101

Anhang III Auflistung der detektierten Sätze mit potenziellen Raumbezug und dessen Ergebnis zur händischen Kodierung 109

Anhang IV Kartendarstellung der einzelnen Überschneidungsbereiche für jedes Airbnb und nach Puffer- oder Isochronen-Distanzen 132

Abbildungsverzeichnis

Abbildung 1 Darstellung eines KDP (Cleve & Lämmel 2016: 6).....7

Abbildung 2 Der Data Mining Prozess (Aggarwal 2015: 4).....10

Abbildung 3 Bestandteile von Textmining (Preethi & Radha 2016: 47).....11

Abbildung 4 Schritte bei der NLP (Dale 2010).....14

Abbildung 5 9-Intersection Matrix nach (Egenhofer, Mark & Herring 1994).....17

Abbildung 6 Erstellung von Geo-text Daten: explizit (l) und implizit (r) (Hu 2018: 5).....18

Abbildung 7 Prozess: geographische Relevanz aus Text (Leidner & Lieberman 2011: 6).....20

Abbildung 8 Eine Mental Map eines Touristen aus Brüssel (Hofmann 2012: 60).....31

Abbildung 9 Isochrone (l) im Vergleich mit Luftliniendistanzen (r) (Wilke 2013).....33

Abbildung 10 Vektordaten: Punkt (a), Linie (b) und Fläche (c) (Paterson & Gregory 2019: 45).....33

Abbildung 11 Beispielhafte Einzugsgebietsanalyse zweier Standorte (grüne und blaue Fläche) und deren Überschneidungsbereich (rote Fläche) (Hintergrund: basemap.at).....35

Abbildung 12 Prozess beim Analysieren von Geo-text Daten (nach Hu 2018: 13).....36

Abbildung 13 Anzahl Airbnbs und Verhältnis von Airbnbs zu Wohnungen je Bezirk.....38

Abbildung 14 Kano-Modell (Hölzing 2008: 85).....42

Abbildung 15 Ablauf der praktischen Analyse.....47

Abbildung 16 Lesen des Codes unter der Verwendung von Pipes.....55

Abbildung 17 Anzahl an Codes je Airbnb.....61

Abbildung 18 Anzahl eindeutiger POIs je Airbnb.....62

Abbildung 19 Unterschied zwischen Puffer und netzwerkbasiertem Einzugsgebiet.....67

Abbildung 20 Beispielhafte Distanzmessung bei Airbnb ID 1 (Hintergrund: basemap.at).....68

Abbildung 21 ideale Erstellung des Überschneidungsbereichs bei Airbnb mit ID 7.....70

Abbildung 22 Screenshot: airbnb.com zu Unterkunft mit ID 4 (Hintergrund: google.com).....71

Abbildung 23 Berücksichtigung der Realnutzungskartierung bei Airbnb ID 4.....73

Abbildung 24 Detektion von möglichen Wohnblöcken durch die Realnutzung.....74

Tabellenverzeichnis

Tabelle 1 Vergleich von strukturierten, semi-strukturierten und unstrukturierten Daten 9

Tabelle 2 Eigenschaften von lokalisierten Objekten und Referenzobjekten.....18

Tabelle 3 Fehlermuster von Tweets mit Standortbezug.....26

Tabelle 4 Allgemeine morphologische Gegebenheiten von Texten verschiedener Korpusse 27

Tabelle 5 Anteil an Sätzen, die durch undefinierbare Grammatik nicht *parsebar* sind 27

Tabelle 6 Familien von räumlichen Beziehungen.....29

Tabelle 7 Beschreibung der verwendeten Daten von Inside Airbnb.....40

Tabelle 8 Liste der Unterkünfte inklusive der Anzahl an Bewertungen 51

Tabelle 9 Gegenüberstellung von R und Python.....52

Tabelle 10 Eingesetzte Pakete und deren wichtigsten Funktionen 55

Tabelle 11 Lexika zu Distanzangaben 56

Tabelle 12 potenziell relevante räumliche und zeitliche Wortfolgen.....57

Tabelle 13 Ergebnis der Messung der Entfernungen 69

Tabelle 14 Lerneffekte nach der praktischen Auswertung.....72

Abkürzungsverzeichnis

GIP	Graphenintegrations-Plattform
GIS	Geoinformationssystem
NER	Named Entity Recognition
NLP	Natural Language Processing
OGD	Open Government Data
POI	Point of Interest
POS	Part-of-speech
IDE	integrated development environment
CLI	command-line interface
RAM	Random-Access Memory
GNU	General Public License

Kurzfassung

Innerhalb von textuellen Daten gibt es ein Potenzial zur Gewinnung räumlicher Information. Dies kann helfen Personen aufgrund deren relativen Entfernungsangaben zu lokalisieren, ohne dass dabei eine umfangreiche Kenntnis der Umgebung vorausgesetzt wird. Um diesen heuristischen Ansatz zu testen, wurden als Basis die Daten von der Onlineplattform Airbnb verwendet. Die Tauglichkeit dieser Quelle liegt darin begründet, da die relative Lage der Unterkunft zu touristischen Sehenswürdigkeiten ein wichtiges Kriterium bei der Buchung ist. Zur Ermittlung von relevanten Tripletten aus dem vorhandenen Korpus wurden Techniken aus dem Textmining und dem Geoparsing eingesetzt. Dies sorgte dafür, dass nur Geo-text Daten für die Analyse übrigblieben, also nur Informationen, die ein Potenzial zur Findung von Positionen aufgrund von quantitativen Entfernungsangaben aufwiesen. Es hat sich gezeigt, dass von den 293 detektierten Sätzen, 20% als relevante Distanzinformationen in ein Geoinformationssystem (GIS) überführt werden konnten. Es wurde auf eine händische Kodierung zurückgegriffen, da in den Sätzen semantische Mehrdeutigkeiten auftraten und das Relatum durch den Computer oftmals nicht zu identifizieren war. Die darauffolgende Analyse hat gezeigt, dass bei zwei relativen Entfernungsangaben keine weitere Bearbeitung notwendig ist und dadurch der Abstand zum Potenzialbereich in dem sich das Airbnb befindet, auf 13,8m verringert wurde. Bei Vorhandensein von mehr als zwei Angaben musste eine visuelle Interpretation angewandt werden, da die Unsicherheiten zu groß waren. Es entstehen mehrere Überschneidungsbereiche, die teilweise nicht automatisch zu reduzieren sind. Nichtsdestotrotz kann sogar in diesen Fällen, bei zusätzlicher Berücksichtigung der Realnutzung der Gebäude, der relevante Häuserblock bestimmt werden.

Abstract

Within textual data there is a potential for obtaining spatial information. This can help to locate people based on their relative distance information, without requiring extensive knowledge of the environment. To test this heuristic approach, the data from the Airbnb online platform was used as the basis. The suitability of this source is since the relative location of the accommodation to tourist attractions is an important criterion when booking. Text mining and geoparsing techniques were used to determine relevant triplets from the existing corpus. This ensured that only geo-text data remained for the analysis, that is, theoretically only information that had the potential to find positions based on quantitative distance information. It has been shown that of the 293 detected sentences, 20% could be transferred as relevant distance information into a geographic information system (GIS). It was based on a manual coding, as in the sentences semantic ambiguities occurred and the relatum was often not identified by the computer. The subsequent analysis showed that with two relative distance indications, no further processing is necessary, thereby reducing the distance to the potential area in which the Airbnb is located to 13.8m. In the presence of more than two indications, a visual interpretation had to be applied because the uncertainties were too great. There are several overlapping areas, some of which cannot be automatically reduced. Nevertheless, in these cases, with additional consideration of the real use of the buildings, even the relevant block of houses can be determined.

1 Einleitung

Das erste Kapitel dieser Arbeit legt fest, welche wissenschaftliche Motivation hinter dem Titel „Möglichkeiten zur Lokalisation von Standorten anhand von quantitativen Entfernungsangaben in Texten mittels Textmining am Beispiel Airbnb“ steht. Darin enthalten ist neben der Problemstellung, die daraus resultierende Relevanz, die Forschungsfragen, deren Beantwortung zur Generierung neuen Wissens beitragen und abschließend der Aufbau der Arbeit.

1.1 Problemstellung

Nutzergenerierte Daten werden als eine dominante Quelle angesehen, wenn es darum geht geographische Gegebenheiten zu bestimmen. Dazu zählen Städte, Straßen oder auch Points of Interest (POI). Der Grund weshalb einer solchen Datengrundlage ein überzeugender Wert zugeordnet wird, ist die Tatsache, dass es sich dabei um lokales Wissen handelt. (Hecht & Gergle 2010) Im Bereich von sozialen Netzwerken wird allerdings angenommen, dass nur in etwa 1-3% aller Beiträge georeferenziert sind. Durch die großen Datenmengen, die in sozialen Netzwerken vorhanden sind, haben sich trotzdem Ansätze entwickelt, wie es möglich ist die Position aus dem Text selbst zu eruieren. (Utomo, Adji & Ardiyanto 2018) Zwei gängige Methoden sind die ‚namentliche Objekterkennung‘, bei der die nutzergenerierten Daten mit einem geographischen Lexikon abgeglichen werden oder der ‚regelbasierter Ansatz‘, mit dessen Hilfe die Orte auf Basis von Mustererkennung innerhalb des Satzes gefunden werden sollen. (Utomo, Adji & Ardiyanto 2018; Karagoz et al. 2016: 197) Das Problem dieser Verfahren ist die grobe Positionierung der Standorte. Es lassen sich POIs meist nicht auf eine einzelne Lage reduzieren, da sich ein potenzieller Aufenthalt meist über mehrere hundert Meter erstreckt. Auf Basis dieser Prämisse wird einer Positionierung von relativen Entfernungsangaben in Texten ein hohes Potential für eine verfeinerte Standortermittlung eingeräumt. (Middleton et al. 2018) Schon die Autoren Leidner & Lieberman 2011 legten das Problem des semantisch-kompositionellen Ansatzes fest. Dieser sagt aus, dass aus einer Entfernungsangabe, einer kardinalen Richtung und einem Standort, ein unscharfer Bereich identifiziert wird, in dem sich das zu referenzierende Objekt befindet. Einige Studien haben sich diesem Problem der relativen Ortsangabe schon angenommen, allerdings in einem kleinmaßstäbigen Bereich, bei dem eine falsche Georeferenzierung schon eine Positionierung in einem anderen Land bedeuten kann. (Kordopatis-Zilos, Papadopoulos & Kompatsiaris 2017; Yuan 2011) Die vorliegende Arbeit hingegen beschränkt sich auf das Wiener Stadtgebiet und sucht Positionen im großmaßstäbigen Bereich auf Basis von relativen Entfernungsangaben. Als Datenquelle dienen Beschreibungen und Bewertungen von Unterkünften, die auf der Onlineplattform *Airbnb* veröffentlicht und von der Internetseite *Inside Airbnb* aufbereitet sind.

1.2 Zielsetzung

Bei der Onlineplattform ‚Airbnb' (ursprünglich: *Airbedandbreakfast*; deutsch: Luftmatratze und Frühstück) handelt es sich um einen Vermittler von Unterkünften für reisende Privatpersonen oder für Geschäftsleute. (Airbnb 2019a) Dabei beruht der Dienst auf dem Prinzip der ‚shared economy' und bedient sich dem Peer-zu-Peer - System. Privatpersonen vermieten demnach direkt an eine andere Person. (Puschmann & Alt 2016)

Neben den Hosts und Einrichtungsmerkmalen hat sich gezeigt, dass auch die Lokalisation der Unterkunft für die Reisenden zu den wichtigsten Attributen für eine vorherige Entscheidungsfindung zählt. (Cheng & Jin 2019) Anbieter haben dabei die Möglichkeit, den genauen Standort zu veröffentlichen oder die Lokalisation zu verschleiern, indem nur ein Radius angegeben wird, innerhalb dessen sich die Herberge befindet. Die Rekonstruktion der exakten Adressen ist dabei im letztgenannten Fall nicht möglich. Eine Positionsschätzung kann in solchen Fällen durch das Konzept des ‚Geoparsing' erreicht werden. Dabei wird festgelegt, welche Wörter in einem Satz Aufschlüsse über die Lokalisation geben könnten und eine Hilfestellung zur automatisierten Verortung bieten. (Zhang & Gelernter 2014) Die zu analysierenden Sätze stammen somit aus den Beschreibungen und Bewertungen der Unterkünfte. Die räumlichen und semantischen Beziehungen werden daraus extrahiert. Nach diesem kontextuellen Übersetzen werden die Relationen in ein Geoinformationssystem (GIS) überführt. Ein Gebiet mit dem potenziellen Standort der Unterkunft kann auf diese Weise ermittelt werden. Das Ziel dieser Thesis ist es nicht die exakte Position von Airbnbs herauszufinden, um diese dann zu veröffentlichen und preiszugeben, sondern weitere Herausforderungen im Forschungsgebiet der Geosemantik im Allgemeinen aufzuzeigen und damit die Überlegungen zur Verfeinerung der Positionsgenauigkeit fortzusetzen. (Middleton et al. 2018)

1.3 Relevanz der Arbeit

Einen Ort anhand von textuellen Quellen zu finden, wird als komplex angesehen, da es auf zwei unterschiedliche Arten erfolgen kann: (1) direkt über den Namen des Ortes, wobei dieser dann auch in einem Gazetteer (geographisches Wörterbuch) georeferenziert vorliegen muss oder (2) über eine relative Angabe der Entfernungen. Letzter Punkt wird als ‚undefinierbare Art' beschrieben. (Moncla et al. 2017) Die vorliegende Arbeit versucht beide Typen zur Identifizierung geographischer Referenz anzuwenden, um den Standort zu finden auf den sich der Text bezieht. In der Literatur ist von der ‚Suche nach einem kontextuellen, geographischen Beweis' die Rede. (Middleton et al. 2018)

Die wissenschaftliche Motivation dieser Arbeit ist dadurch begründet, da zivile Sicherheitseinrichtungen, wie die Polizei, Rettung, Feuerwehr oder das Bundesheer in Krisensituationen auf Informationen aus sozialen Netzwerken zurückgreifen. Daraus lässt sich meist ablesen an welchem Ort sich ein Unheil ereignet hat. (Earle, Bowden & Guy 2011) Es gibt keine vergleichbaren Studien, die sich mit der Standortermittlung durch relative Entfernungsangaben aus soliden und ausreichend verfügbaren Texten

beschäftigen. Der theoretische Ansatz quantitative Informationen einzusetzen, um räumliche Gegebenheiten aus textuellen Daten zu ermitteln ist nicht neu und wurde bereits von Papadias & Sellis 1994 angedacht. Allerdings überwogen damals noch die folgenden Nachteile:

- Die Genauigkeit von quantitativen Repräsentationen ist meist nicht genügend
- In räumliche Prozesse fließen meist qualitative anstatt quantitativer Informationen
- Qualitative Informationen sind meist günstiger

Nichtsdestotrotz wird, wie oben erwähnt, einer solchen Möglichkeit, ein hohes Potential vorhergesagt. Als Grund dafür, wieso es bis heute wenig Forschung in diese Richtung gegeben hat, ist die Tatsache, dass vorwiegend Twitterdaten als Grundlage verwendet werden. In einem solchen Fall sind relative Entfernungsangaben eher die Seltenheit. Außerdem haben aktuelle Studien meist eine Echtzeit-Geocodierung als Ziel, ohne dass der Mensch aktiv in den Prozess eingreifen muss. (Utomo, Adji & Ardiyanto 2018) Bei den Bewertungen und Beschreibungen von Airbnb-Unterkünften werden Standorte teilweise mit relativen Entfernungen angegeben und bieten aus diesem Grund eine Chance neue Lokalisations-Prozesse auf Basis von nutzergenerierten Daten zu entwickeln.

1.4 Forschungsfragen

Aufgrund der vorliegenden Daten von Inside Airbnb soll eine Positionsschätzung aus relativen Entfernungsanagaben durchgeführt werden. Der Fokus der Arbeit liegt darauf, dass mittels Textmining die räumlich relevanten Sätze gefunden werden und folglich die POIs, sowie die dazugehörige Entfernungsangabe zu extrahieren. Es soll auf Basis von zwölf exemplarischen Standorten ein heuristischer Ansatz zur Verfeinerung der Positionsgenauigkeit auf Basis von nutzergenerierten Daten erstellt werden.

Zur Beschreibung dieses Forschungsgegenstandes benötigt es einer Forschungsfrage und dazugehörigen Arbeitsfragen. Es gilt vorwiegend zu klären, ob es möglich ist einen Standort unter Verwendung von relativen Entfernungsangaben zu ermitteln. Die daraus abgeleitete Forschungsfrage lautet:

- *Ist es möglich den Standort auf Basis von relativen Entfernungsangaben, innerhalb von nutzergenerierten textuellen Daten, zu bestimmen?*

Zur Klärung der obigen Forschungsfrage werden drei Arbeitsfragen herangezogen, um auch potenziell neue Ansätze zu evaluieren. In einem ersten Schritt soll festgestellt werden, inwieweit Textmining zur Geolokalisation bereits erforscht ist, welche Ansätze davon übernommen werden können und welche Mindestanforderungen ein Satz haben muss, um für die Standortermittlung von Relevanz zu sein.

- *Wo befindet sich der Forschungsstand zum Thema ‚Geolokalisation auf Basis von Textmining' und inwieweit können daraus folgende Erkenntnisse unterstützend auf diese Arbeit wirken?*

- *Welche Hypothesen sind bei der Extraktion von räumlich relevanten Inhalten aus textuellen Informationen festzulegen, um eine sinnvolle Datenbasis für die Analyse im GIS zu erstellen?*

Im Rahmen der späteren Analyse ist anzunehmen, dass es aufgrund der semi-strukturierten Datenquelle, dem heuristischen Ansatz bei der Extraktion von räumlicher Information und der damit einhergehenden Mehrdeutigkeit von Ausdrücken, zu Herausforderungen bei der Auswertung kommt. Deshalb widmet sich eine Arbeitsfrage, welche Probleme bei der praktischen Analyse aufgetreten sind.

- *Welche praktischen Problemstellungen ergeben sich bei der Aufbereitung und Verarbeitung von textuellen Daten mit Raumbezug für ein Geoinformationssystem?*

1.5 Verwendete Methodik und Aufbau der Arbeit

Die Masterarbeit besteht aus einem theoretischen und einem praktischen Teil. In der Theorie ist der derzeitige Forschungsstand auf Basis einer Literaturanalyse wiedergegeben und der Einfluss auf die vorliegende Arbeit bestimmbar. Im praktischen Teil werden die Daten von Inside Airbnb herangezogen. Diese stammen ursprünglich von der Internetseite Airbnb. Auf der Webseite von Inside Airbnb werden alle für den Nutzer sichtbaren Daten von ausgewählten und bevölkerungsreichen Städten in einer menschen- und maschinenlesbaren Form angeboten (*.csv*-Dateien). Die Daten werden in unregelmäßigen Abständen, aber doch mehrmals jährlich aktualisiert. (Cox 2018a) Die vorhandene Güte und Objektivität der Quelle wird durch den Verweis auf mehreren Autoren belegt. (Cheng & Jin 2019; Ma et al. 2017; Guttentag 2016) Für die Analyse werden nur jene Unterkünfte herangezogen, welche den tatsächlichen Standort der Unterkunft auf der Internetseite von Airbnb veröffentlicht haben. Mit dieser Vorgehensweise ist eine ethische Grenze gezogen, die garantieren soll, dass keine Standorte von Unterkünften gefunden werden, bei denen es die Gastgeber nicht möchten.

Im praktischen Teil, bei dem es darum geht, die räumlich relevanten Informationen aus dem Text zu extrahieren, wird immer auf den gleichen Ablauf zurückgegriffen, der ‚namentlichen Objekterkennung'. Bei dieser Methodik wird ein eigenes geographisches Lexikon angelegt und in diesem Fall mit den zerlegten Sätzen aus den vorhandenen Airbnb-Daten abgeglichen. Als Erweiterung zu dieser Methode wird ein ‚regelbasierter Ansatz' angewandt. In diesem Fall wird zusätzlich zur Toponymsuche auch eine Mustererkennung eingesetzt. (Utomo, Adji & Ardiyanto 2018; Karagoz et al. 2016: 197) Die zu analysierenden Sätze sind bereits durch ein eigens erstelltes Skript vorgefiltert, denn diese stellen nur noch Informationen dar, die auch räumliche oder zeitliche Angaben enthalten. Diese Ansätze stellen allesamt Methoden des Textminings dar, mit einem Schwerpunkt auf räumliche Fragestellungen. Dieser Teil soll praktisch mit dem Statistik-Tool *R* und den darin angebotenen Bibliotheken gelöst werden.

Nachdem die Datenaufbereitung und -analyse soweit erfolgt ist, wird in ein GIS übergegangen. Die räumlichen Ausgangsdaten sind die offiziellen POIs, die von der Stadt Wien zur Verfügung gestellt werden. Diese beinhalten sowohl „Sightseeing"-Objekte als

auch Haltestationen des öffentlichen Verkehrs. Die relativen, räumlichen Distanzangaben der Benutzer werden als gleichmäßiger Puffer rund um den POI und als Wegstrecke auf den österreichweit verfügbaren, behördlichen Verkehrsgraphen abgebildet (Graphenintegrations-Plattform, GIP). Es wird angenommen, dass die Entfernungsangaben der Gäste zu den berücksichtigten POIs im Text für Fußgänger gelten. Die Überschneidungspunkte der daraus entstehenden Flächen sollen eine wahrscheinliche Lokalisation der Unterkunft darstellen. Bei einer Verschleierung der Position von Standorten verwendet Airbnb einen potenziellen Bereich mit einem Durchmesser eines Kilometers. Innerhalb dieses Korridors befindet sich die Unterkunft. Das bedeutet, dass eine Verkleinerung des Suchradius auf unter einem Kilometer, die verwendete Methodik und deren Sinnhaftigkeit bestätigen würde. Der Unterschied zwischen errechnetem und tatsächlichem Standort soll des Weiteren gemessen und näher untersucht werden. Im Rahmen der Auswertung, werden sich Empfehlungen ergeben, die in weiteren Forschungen zu berücksichtigen sind.

Die Arbeit gliedert sich in 7 Überkapitel. In Abschnitt 2 wird der aktuelle Forschungsstand zu unterschiedlichen, thematisch relevanten Inhalten festgehalten. Dazu zählen das Data Mining, die natürliche Sprachprozessierung oder das Geoparsing. Im Allgemeinen widmet sich das Kapitel der Herausforderung, wie Raum und Text miteinander verknüpft werden können.

Das Kapitel 3 zeigt die Datenquellen auf, die in dieser Arbeit zum Einsatz kommen. Es handelt sich um freiverfügbare Informationen aus dem Internet. Dahingehend wird erläutert wieso es Motivationen gibt, solche Daten mit anderen Menschen zu teilen.

Das Set-Up der Arbeit ist in Kapitel 4 dargestellt. Es soll die Entscheidungen für einzelne Technologien nachvollziehbar machen und den angewandten Untersuchungsraum näher umschreiben. Außerdem wird ein erster Schritt in der praktischen Verarbeitung getätigt, in dem bereits räumlich relevante Information aus den textuellen Daten extrahiert und klassifiziert wird.

Kapitel 5 bildet einen Übergang zu den eingesetzten GI-Systemen. Bevor dieser Schritt vollzogen werden kann, sind bestimmte Vorarbeiten zu leisten, wie das Wählen der Projektion oder die Klärung der treffendsten Distanzangaben bei einer Mehrfachnennung eines POIs. Die daraus entstehenden Ergebnisse werden visualisiert und vermessen, um einen Vergleich zu ermöglichen.

Die Beantwortung der oben definierten Forschungsfragen befindet sich in Kapitel 6. Obendrein sind noch Handlungsempfehlungen für die zukünftigen Analysen von relativen Entfernungsangaben, die sich durch Problemstellungen aus der praktischen Arbeit ergeben haben, enthalten.

Den Abschluss bildet Kapitel 7. Dieses enthält eine Zusammenfassung der gesamten Arbeit und eine wissenschaftlich fundierte Aufarbeitung von zukünftigen Forschungsthematiken in diesem Bereich.

2 Stand der Forschung

In Abschnitt 2 soll der derzeitige Stand der Forschung und somit das theoretische Grundgerüst dargelegt werden. Das ist zum einen das Data Mining, welches dem Text Mining und damit der natürlichen Sprachprozessierung übergeordnet ist. Diese wird zur Deutung von Text, zum Beispiel zur Festlegung ob überhaupt räumliche Information enthalten ist, eingesetzt. In diesen Kontext fällt auch das Geoparsing und alle dazugehörigen räumlich relevanten Begrifflichkeiten Des Weiteren wird das Konzept von Geo-text Daten nach Hu 2018 vorgestellt. Den Abschluss bilden die Unterkapitel zur menschlichen Wahrnehmung von Distanzen und die Methoden zur Berechnung von Vektorpuffern und Isochronen.

2.1 Data Mining

Dieser Abschnitt soll darlegen, wieso es Data Mining gibt und mit welchen verschiedenen Datentypen im Laufe einer Analysetätigkeit zu rechnen ist. Außerdem sollen die grundlegenden Begrifflichkeiten von Text Mining und dem räumlichen Data Mining dargelegt werden. Dies bildet die theoretische Basis für die praktischen Überlegungen zur Verknüpfung von Raum und Text.

2.1.1 Prinzipien

Data Mining (DM) ist die Wissenschaft von der Sammlung, Bereinigung, Prozessierung, Analyse und Erkenntnisgewinnung von und aus Daten. Je nach Problemstellung, Applikation und Datenrepräsentation ändert sich die Definition, die Data Mining am treffendsten beschreibt. Die Motivation für den Einsatz solcher Techniken ist, dass die meisten Datenquellen unstrukturiert und unübersichtlich sind. Außerdem hat es in den letzten Jahren einen explosionsartigen Anstieg an Datenmengen gegeben. (Aggarwal 2015: 1) Zurzeit werden in etwa 33 Zettabytes an Daten pro Jahr weltweit gesammelt. In den nächsten sieben Jahren wird dieser Wert um 61% auf 175 Zettabytes pro Jahr ansteigen. (Cooney 2018) Diese Datenmengen werden beispielsweise durch die Geldtransaktionen, die Fernerkundung, das Internet oder den Verkehr gesammelt. Es lassen sich ohne die Anwendung von Data Mining keine sinnvollen Ableitungen zu zukünftigen Verhalten feststellen. (Aggarwal 2015: 1)

Als die Wurzel des Data Minings wird die Statistik angesehen. In den frühen 1980er Jahren wurde in diesem Zusammenhang von der explorativen Datenanalyse gesprochen. Diese konnte auf zwei verschiedene Arten umgesetzt werden: (Gorunescu 2011: 2–4)

- Berechnungsmethoden: Beginnend bei der deskriptiven Statistik, wie Median, Durchschnitt, Korrelationen, Häufigkeitstabellen, bis hin zu Klassifikationsbäumen oder Regressionsmodellen (linear oder nicht-linear)

- Visualisierung: Sie wird als die mächtigste Methode der Datenexploration und gleichzeitig attraktivste Methode angesehen, da Histogramme, Verteilungen oder Boxplots relativ einfach und von einem Großteil an Beteiligten analysiert und interpretiert werde kann.

Das Data Mining an sich ist nur ein Teil des Wissensentdeckungsprozesses (engl. knowledge discovery process, KDP). Innerhalb eines KDP können beispielsweise mehrere DM-Prozesse sequenziell durchgeführt werden. Die Verwendung eines KDP ist darin begründet, da meist mehrere Iterationen zwischen Analysten und Datenbesitzer notwendig sind, um einen Konsens zu erzeugen. Ein solcher Prozess steuert wie die Datenextraktion vonstattengeht, in welchem Format und wie die Daten gespeichert werden, welche Algorithmen darauf zugreifen, wie die Skalierbarkeit erreicht wird und zum Schluss wie die Ergebnisse visuell darzustellen sind. (Cios et al. 2007: 9–11) Ein solcher KDP ist in Abbildung 1 dargestellt. Darin werden in einem ersten Schritt nur die, für die Analyse relevanten Daten selektiert und in eine Datenbank oder eine andere Verarbeitungsumgebung eingefügt. Die Datenvorverarbeitung behandelt fehlende oder fehlerhafte Datenpunkte. Bei der Transformation erfolgt die Umwandlung in adäquate Datenformate – beispielsweise eine Faktorisierung von nummerischen Werten. Im Data Mining wird dann nach Mustern gesucht. Die daraus resultierenden Ergebnisse werden letztlich evaluiert und interpretiert. (Cleve & Lämmel 2016: 5–6)

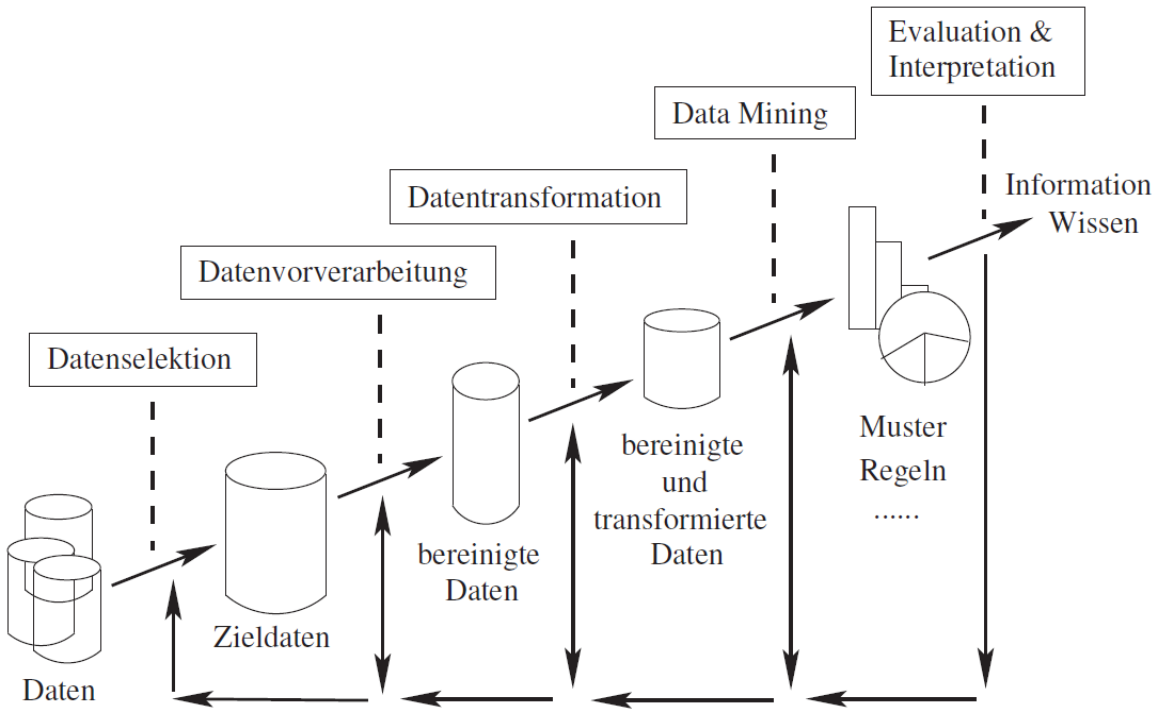


Abbildung 1 Darstellung eines KDP (Cleve & Lämmel 2016: 6)

Als andere Darstellungsmöglichkeit ist noch das CRISP-Modell (engl. **Cr**oss **I**ndustry **S**tandard **P**rocess for Data Mining) zu erwähnen. Dieser Prozess ist allerdings mehr auf die Bedürfnisse von Firmen bzw. der Industrie getrimmt. Darin enthalten sind beispielsweise drei Schritte, die in Abbildung 1 nicht vorkommen. Das sind die Abschnitte zum

„*Verstehen der Aufgabe*“ (Klären der vorhandenen Ressourcen, Ausgangssituation oder beachten der Unternehmensziele), zum „*Verständnis der Daten*“ (zunächst wird das Ziel definiert und danach werden die benötigten Daten angefordert – deren Nutzen ist auch zu dokumentieren) und zu guter Letzt zum „*Einsatz im und Konsequenzen für das Unternehmen*“. Bei einem erfolgreichen DM-Projekt wird das Ergebnis mit in den Regelbetrieb aufgenommen. (Cleve & Lämmel 2016: 8–9)

Unterschied zwischen strukturierten, semi-strukturierten und unstrukturierten Datentypen

Je nachdem wie Daten generiert und gespeichert werden, lassen sich unterschiedliche Typen festlegen. Diese lassen sich in drei Typen unterscheiden: (Die Unterschiede sind zusammenfassend in Tabelle 1 aufgelistet)

- Strukturierte Daten

  Die Daten lassen sich einer Struktur oder einem Schema zuordnen und sind meist in einer Datenbank abzuspeichern. Die Information ist in Spalten und Zeilen enthalten. Es stellt eine Art dar, die es ermöglicht, dass daraus direkt Information gewonnen wird, ohne die Daten selbst zu verändern. Sowohl der Computer als auch der Mensch kann die Daten durch die effiziente Ordnung, „verstehen“. (Monica & Kumar 2013; Bey 2014)

- Semi-strukturierte Daten

  Es handelt sich weder um Rohdaten noch um Daten, die in herkömmlichen Datenbanksystemen gespeichert werden. Aus diesem Grund ist auch nicht notwendig, dass die Daten einem vordefinierten Schema folgen. Nichtsdestotrotz können semantische Unterschiede sichtbar gemacht werden, wodurch Hierarchien erkennbar sind. Außerdem kann postpriori auf Basis der bereits bekannten Instanzen ein Schema definiert werden. Die meisten Texte in der Informationstechnologie sind als semistrukturiert einzustufen. (Sint et al. 2009; Monica & Kumar 2013)

- Unstrukturierte Daten

  Bei diesem Datentyp werden zwei Arten unterschieden. Es sind entweder nicht-textuelle Daten, wie Videos, Bilder oder Musik oder textuelle Daten, wie E-Mails, Memos oder Kurznachrichten. Die Daten könne nicht in Zeilen und Spalten in ein relationales Datenbanksystem gespeichert werden. Es gibt keine Notwendigkeit ein Schema zu definieren, da die Daten sehr flexibel aufgebaut sind. Dieser Vorteil führt allerdings dazu, dass keine kontrollierte Navigation durch die Daten möglich ist. (Sint et al. 2009)

	strukturierte Daten	**semi-strukturierte Daten**	**unstrukturierte Daten**
Technologie	relationale Datenbank-Tabelle	XML, JSON	Zeichen oder binäre Daten
Transaktions-Management	gereiftes Transaktions-Management	adaptiertes Transaktions-Management vom RDBMS, aber nicht ausgereift	Kein Transaktions-Management
Versionsmanagement	Versionierung über Tupel, Reihen, Tabellen, etc.	Versionierung über Tupel möglich	nur als Ganzes zu versionieren
Flexibilität	strikt abhängig vom Schema	flexibel und Schema tolerant	sehr flexibel, kein Schema vorhanden
Skalierbarkeit	DB-Schema Skalierung ist schwierig	DB-Schema Skalierung ist einfach	hoch skalierbar
Robustheit	sehr robust	Neue Technologie - nicht weit verbreitet	-
Abfrage-Leistung	strukturierte Abfragen und komplexe Joins möglich	Abfragen über anonyme Knoten sind möglich	nur textuelle Abfragen

Quelle: (Monica & Kumar 2013: 429)

Tabelle 1 Vergleich von strukturierten, semi-strukturierten und unstrukturierten Daten

2.1.2 Vorgehensweise

Der Data Mining Prozess folgt immer einem bestimmten Ablauf. (Abbildung 2) An erster Stelle steht die Datenbasis, die mittels einer Umfrage, auf Basis von Sensordaten oder mittels abgreifen von Informationen aus dem Internet erstellt wurde. Die Daten sind dann meist in einer Datenbank abgelegt, wo sie dann für die weitere Verarbeitung bereitstehen. Ist dieser Schritt abgeschlossen, erfolgt die Daten-präprozessierung, denn die Daten sind meist unstrukturiert und mannigfaltig. Es gilt durch eine *Feature*-Extraktion nur jene Variablen zu selektieren, die für eine weitere Analyse von Interesse sein könnten bzw. es auch sind. Das dient vor allem dazu, um die Datenmenge zu reduzieren und einen Überblick zu ermöglichen. Innerhalb der übrig geblieben Informationen kann es zu Inkonsistenzen oder auch fehlenden Einträgen kommen. Solche Fälle sind zu entfernen oder auch mit Berechnungen zu füllen. Da jede Data Mining Applikation einmalig ist, stellt die analytische Prozessierung die größte Herausforderung in diesem Ablauf dar (unterschiedliche Datentypen, statistische Methoden, Wissenshintergrund der handelnden

Personen). In diesem Schritt können auch Rückkopplungen zu beiden vorherigen Blöcken durchgeführt werden. (Aggarwal 2015: 4–6)

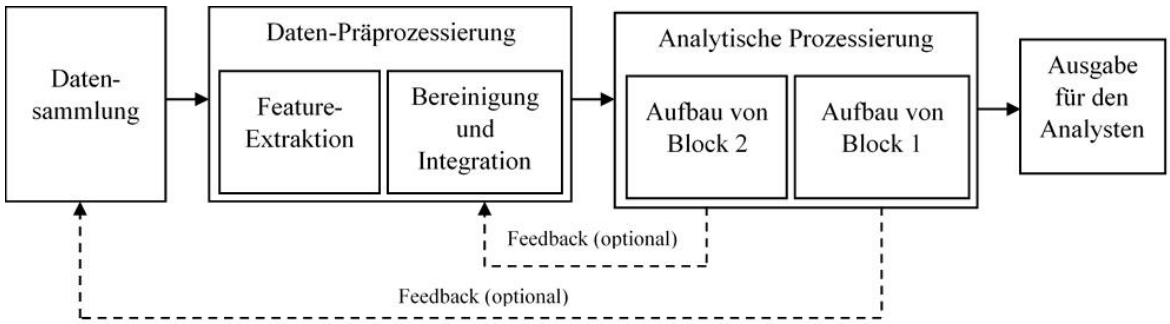


Abbildung 2 Der Data Mining Prozess (Aggarwal 2015: 4)

Zu den oben erwähnten verschiedenen Datentypen, stellt neben den quantitativen, binären, textuellen oder zeitlichen Datentypen auch der Raum eine eigene Datenstruktur dar. Durch die Berücksichtigung der Ausdehnung, als ein Punkt, eine Linie oder eine Fläche wird dieser Typ als komplexer angesehen. Solch räumliche Daten enthalten aber auch beinahe immer nicht-räumliche Informationen. Als Beispiele sind hier Arbeitslosenraten oder das Bruttoinlandsprodukt zu erwähnen. Diese Daten können mit klassischen Data Mining Techniken verarbeitet werden. Außerdem ist die Beziehung der einzelnen Objekte nicht explizit, sondern meist implizit. Die Daten sind somit voneinander nicht unabhängig. Zum Beispiel ist eine Ordnung der Objekte nach rein räumlichen Informationen deshalb nicht möglich. Die räumlichen Attribute eines räumlichen Objektes stehen meist in Beziehung zu einer Position (Längen-, Breitengrad oder Höheninformationen). (Shekhar, Zhang & Huang 2005)

2.1.3 Textmining

Im Vergleich zu herkömmlichen Daten, die meist in Datenbanken sind, sind textuelle Daten unstrukturiert, unsicher einzuordnen und schwierig zu verarbeiten. Nichtsdestotrotz sind Texte der herkömmlichste Weg, Informationen zu transferieren. Beim Textmining, das seine Anfänge in den 1980er Jahre hatte, werden Techniken aus der Mathematik, Statistik, Linguistik und Mustererkennung eingesetzt. Es setzt sich aus vier verschiedenen Bestandteilen zusammen, die allesamt als gleichwertig anzusehen sind. (Abbildung 3) Das Ziel ist immer, aus unstrukturierten textuellen Daten, strukturierte Informationen zu erhalten. (Preethi & Radha 2016)

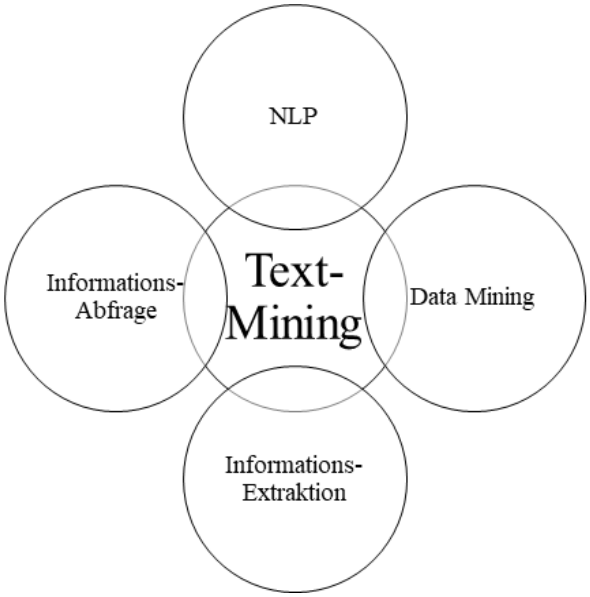


Abbildung 3 Bestandteile von Textmining (Preethi & Radha 2016: 47)

- Informationsabfrage

  Obwohl Textdaten als unstrukturiert anzusehen sind, müssen die Daten abgespeichert sein. Das Speichersystem muss eine Suche nach bestimmten Wörtern oder Informationen zulassen und die Daten für weitere Analysen, wie die NLP, exportieren können. (Miner et al. 2012) Der Export kann auch ein einstellbarer Filter sein, der nur relevante Informationen durchlässt. (Preethi & Radha 2016)

- Informations-Extraktion

  Bei der automatisierten und strukturierten Extraktion von Information aus textuellen Daten wird im Textmining von Informations-Extraktion gesprochen. In den meisten Fällen wird versucht die Ausgaben der NLP zu deuten. Die Informations-Extraktion stellt den Prozess dar, strukturierte Information aus unstrukturierten oder semi-strukturierten Daten zu gewinnen. (Miner et al. 2012; Preethi & Radha 2016)

Die Erläuterung von NLP, die auch in Abbildung 3 dargestellt ist, erfolgt dann in Abschnitt 2.2. Die Grundlagen zum Data Mining sind bereits im vorherigen Kapitel 2.1.1 niedergeschrieben worden.

2.1.4 Räumliches Data Mining

Der Data Mining Prozess ist für den räumlichen Datentyp besonders interessant, da solche Daten die Basis für Big Data darstellen. Ohne den Grundprinzipien des Data Mining wäre es nicht möglich großes, möglicherweise auch verteiltes Datenmaterial zu verarbeiten. Der Begriff „Big Data“ wird immer dann verwendet, wenn Standard Hard- und Software nicht

mehr ausreicht, um die vorhandenen Datenmengen zu verarbeiten. (Bhadani & Jothimani 2017)

Beim räumlichen Data Mining gilt es geometrische und statistische Beziehungen innerhalb von räumlichen Daten sichtbar zu machen. Um effizient das Hauptziel, der Gewinnung von neuer Information zu erreichen, ist das Vorhandensein von nicht-räumlichen Information unabdinglich. (Perumal et al. 2015) Räumliche Daten sind multidimensional und autokorreliert, denn diese haben eine Lokalisation, eine Ausrichtung, eine Fläche und eine Orientierung. Attributwerte hingegen sind unabhängig von geometrischen Gegebenheiten. Räumliches Data Mining ist eine Erweiterung der klassischen Data Mining Techniken und benötigt eine Kombination von räumlichen Daten und den dazugehörigen Attributwerten. (McKinney & Cai 2002) Es werden nach Perumal et al. 2015 vier unterschiedliche Aufgaben unterschieden, die zum Beispiel mit den Softwares ArcGIS oder GeoMiner oder den Programmiersprachen R und Python gelöst werden können:

- Räumliche Klassifikation

  Je nachdem welche Attribute ein räumliches Objekt aufweist, wird es einer bestimmten Klasse zugewiesen. (Brinkhoff & Kriegel 1994) Solche Techniken werden beispielsweise in der Fernerkundung eingesetzt, wenn bestimmte Pixelwerte eines Rasters in naturräumliche Gegebenheiten eingeteilt werden. In einem solchen Fall sind durch ein festgelegtes Regelwerk auch die Nachbarschaftsbeziehungen der einzelnen Pixel zueinander relevant. (Perumal et al. 2015)

- Räumliche Assoziations-Regeln

  Innerhalb der Daten gilt es Assoziationen, Muster oder Korrelationen in Form von Regelmäßigkeiten zu detektieren. Solche Regeln beschreiben eine Menge an Features, die wiederum eine andere Menge an Features in der räumlichen Datenbank beschreiben. (Lee et al. 2007)

- Räumliches Clustering

  Das räumliche Clustering basiert auf Distanz- oder Richtungsbeziehungen. Alle Elemente innerhalb eines Clusters weisen eine hohe Ähnlichkeit auf und sind von einem anderen Cluster stark verschieden. Diese Ähnlichkeit kann mittels Distanz gemessen werden. (Perumal et al. 2015)

- Trenderkennung

  Bei dieser Technik wird beobachtet, wie sich die nicht-räumlichen Attribute ändern, während sich die räumliche Startposition verändert hat. Auch hier wird versucht eine Mustererkennung durchzuführen, allerdings mit einem Augenmerk auf die räumliche Veränderung. Beispiele für das Anwendungsgebiet sind der Transport, der Tourismus oder die Landwirtschaft. (Perumal et al. 2015)

2.2 Natürliche Sprachprozessierung

In der ersten Hälfte des 21. Jahrhundert stimmte die Mehrheit der Philosophen überein, dass für ein Verständnis von Sprache eines Senders, der Empfänger über eine ebenso gute kontextuelle Wahrnehmung verfügen muss. Dies änderte sich, als zu Beginn der zweiten Hälfte des 21. Jahrhunderts der Linguist Noam Chomsky das Prinzip der „Universal-Grammatik" feststellte. Er fand heraus, dass alle Sprachen denselben grammatikalischen Prinzipien folgen. Diese Theorie festigte er mit der These, dass alle Kinder auf der Welt deren Muttersprache oder auch Erstsprache auf dieselbe Art und Weise erlernen. (Chomsky 1956)

Durch die Möglichkeit der Festlegung von Regeln für alle Sprachen, konnten diese auch dem Computer beigebracht werden. In der Informationstechnologie wird die computergestützte Analyse von digitalen Text- oder Audiodateien als natürliche Sprachprozessierung (engl. *Natural Language Processing*, NLP) geführt. Die NLP ist ein multidisziplinäres Arbeitsgebiet, dass sich aus Computerwissenschaften, künstlicher Intelligenz und Computersprache zusammensetzt. Dieses Forschungsgebiet ist durch die Mehrdeutigkeit von natürlichen Sprachen entstanden, denn im Gegensatz zur formalen Sprache (z.B.: Formeln oder Programmiersprachen) folgt diese keinen Regeln. (Moncla 2016)

Die Aussage von Moncla 2016, dass die natürliche Sprache doch keinen Regeln folgt, steht in einem Spannungsverhältnis zu den Thesen von Chomsky 1956. Dies hat sich aber vor allem durch das digitale Zeitalter und die Nutzung von sozialen Netzwerken, als Kommunikationsplattformen, rasant entwickelt. Der geschriebene Satz oder Text eines Benutzers in einem Blog, Tweet oder ähnlichem wird als eine eigene Sprache angesehen. Diese folgt wiederum eigenen Regeln, die mit jenen aus dem Sprachunterricht in den Schulen nicht zu vergleichen sind. Daraus folgt schließlich, dass es das Ziel der NLP ist, dem Computer trotz alledem ein natürliches Sprachverständnis zu lehren, um Mehrdeutigkeiten und Unsicherheiten in der Interpretation zu vermeiden. (Martinez 2010)

Die natürliche Sprachprozessierung wird traditionell in einzelne Stufen zerlegt. Ein wichtiger Bestandteil ist die generelle Betrachtungsweise eines Satzes. In der dazugehörigen Theorie wird zwischen Syntax, Semantik und Pragmatik innerhalb eines Satzes unterschieden. Die syntaktische Analyse ergibt eine Reihung und Struktur der Wörter, welche wiederum eine Basis für die symbolische Deutung (Semantik) darstellt. Eine abschließende Feststellung der Pragmatik hilft die Bedeutung der Aussage eines Satzes zu analysieren. In der Praxis wird neben den drei oben genannten Analysen noch eine Vorprozessierung benötigt. Der gesamte Prozess ist in Abbildung 4 dargestellt. (Dale 2010)

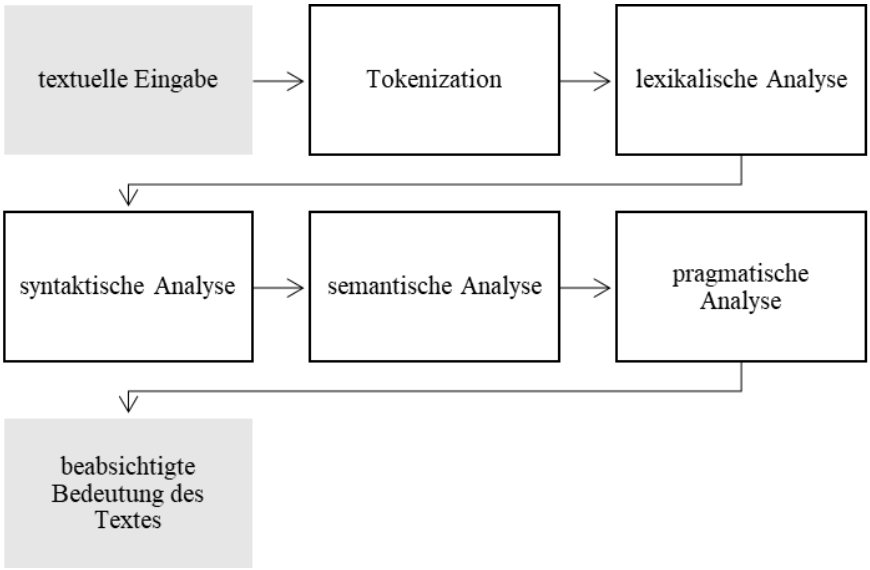


Abbildung 4 Schritte bei der NLP (Dale 2010)

Zur Vorprozessierung zählen die sogenannte Tokenisierung, sowie die lexikalische Analyse. Bei der Tokenisierung ist nur die Wort- oder Satzsegmentierung von Interesse. Ein Satz wird in einzelne Zeichenketten getrennt, damit die Grenzen von Wörtern gefunden werde können – jeder Satz wird danach in simple Einheiten, sogenannte Token, zerlegt. Ein Token kann Punkte, Sonderzeichen, Nummern, Wörter oder gesamte Sätze darstellen. Je höher die Qualität eines „Tokeniser" ist, desto vertrauenswürdiger sind die Ergebnisse, die aus dem NLP-Prozess hervorgehen. Je nach Anwendung sind dafür Regeln zu definieren. Wenn zum Beispiel die Maßeinheit bei einer erkannten Zahl wichtig und auch zugehörig ist, so ist dies festzulegen. Des Weiteren wird im angloamerikanischen Raum der Punkt als Dezimaltrennzeichen und der Beistrich als Tausendertrennzeichen eingesetzt, während es im deutschsprachigen Raum umgekehrt ist. Es muss auch generell ein klar ersichtliches Trennzeichen für die Token vorliegen. Mögliche Problemstellungen sind das nicht Vorhandensein von Leerzeichen in der Sprache. (z.B.: chinesisch oder koreanisch) Als weitere Herausforderung gilt eine unterschiedliche Kodierung (ASCII, Unicode, etc.) von Texten. Dies ist eine exemplarische Auflistung von Problemen, die beim Einsatz eines „Tokenisers" auftreten können und während des Prozesses, mittels einer Stichprobe, zu überprüfen sind. (Palmer 2010; Jackson & Moulinier 2002) Für diese Arbeit sollten die angegeben Fehlerquellen kein Grund für das Scheitern der Analyse darstellen, da nur auf eine Textquelle zurückgegriffen wird und in deutscher, als auch englischer Sprache, die Wörter durch ein Leerzeichen und Sätze mittels Punktation getrennt sind.

Bei der lexikalischen Analyse wird die Ausprägung (in der Linguistik: Morphologie) und die Beziehung von Wörtern zueinander untersucht. Anwendungsbeispiele sind die Suche nach dem Wortstamm oder die Wahrnehmung von unterschiedlichen, zeitlichen Ausprägungen eines Wortes. (Dale 2010) Für diesen Schritt außerdem von Wichtigkeit ist die Beachtung der lexikalischen Mehrdeutigkeit oder in der Fachsprache, die Ambiguität. So kann das Wort „Bank", für ein Finanzinstitut, eine Sitzmöglichkeit oder für den Teil eines Flussufers stehen. Menschen können dies ohne Probleme unterscheiden, da sie den

Kontext verstehen, als auch über ein Wissen zur realen Welt verfügen – im Gegensatz zum Computer. (Jackson & Moulinier 2002) In dieser Arbeit wird dieser Schritt mittels Anwendung eines eigens erstellten Gazetteers erleichtert. Für vertiefende Fortbildung in diesem Bereich ist ansonsten eine gute Kenntnis der Kombinatorik unerlässlich.

Die eigentliche Prozessierung selbst ist besonders dann eine Herausforderung, wenn kein vorheriges Wissen über den Text vorhanden ist. Die Untersuchung der Syntax eines Satzes ist, für einen Algorithmus nichts anderes als die Bestimmung der Wortarten und Wortgruppen nach drei Elementen. (Maggini 2010)

- Wesentliche Bestandteile

  Ein in dessen atomare Token zerlegter Satz, wird in Gruppen von Wörtern zusammengefasst. Als Beispiel gilt hier die Nominalphrase, die aus Nomen, Adjektiven und Konjunktionen besteht – *der alte Mann* oder *ein großer und roter Teppich.*

- Grammatikalische Beziehungen

  Repräsentiert die Formalisierung einer Satzstruktur, als Verbindung zwischen Subjekt und Objekt. *Er* [Subjekt] nahm *den großen Hammer* [Objekt].

- Subkategorisierung und Abhängigkeiten

  Es gibt bestimmte Regeln, die Einschränkungen zwischen Wörtern und phrasalen Gruppen wiedergeben. So kann beim Verb *finden* immer nur eine Nominalphrase vorrangegangen sein (in der englischen Sprache ist dies umgekehrt).

Eine syntaktische Analyse bei nutzergenerierten Daten ist schwierig, da in diesem Zusammenhang meist grammatikalische Regeln für einen Abgleich bestimmend sind, die aber hier nicht angenommen werden können, denn wie oben beschrieben folgen nutzergenierte Informationen keinen Regeln. Außerdem finden sich Abkürzungen von Nomen oder Rechtschreibfehler, die eine Zerlegung in die wesentlichen Bestandteile oder die Testung auf grammatikalische Beziehungen nicht ermöglichen. (Ljunglöf & Wiren 2010) Zur Bestimmung der Semantik wird versucht einzelne Wörter zu extrahieren und abzuleiten, wie eine mögliche Bedeutung des Satzes aussehen könnte. Dieser Schritt ist bei Dale 2010 abgeschwächter, da bereits eine lexikalische Analyse erfolgt. Diese soll bereits Homonyme (ein Wort - mehrere Bedeutungen: s. oben: *Bank*) oder Synonyme (anderes Wort – selbe Bedeutung: *Geschenk* oder *Präsent*) erkennen. Anhand der Vorprozessierung ist es somit möglich die Semantik des Satzes zu bestimmen. Beispiele hierfür wären Texte zu Todesanzeigen oder auch Nachrichten zu politischen Themen. (Goddard & Schalley 2010)

Zusätzlich zu den obigen Theorien zur Syntax, Semantik und Pragmatik gibt es noch weitere Begrifflichkeiten, die im Zusammenhang mit der NLP stehen. Ein *Korpus* ist eine Menge an textuellen Daten. Die eindeutigen Wörter in einem Korpus werden in einem *Lexikon* abgespeichert. Für weitere allgemeine Information zur NLP sei auf Martinez 2010 verwiesen.

Sobald eine Datei vorliegt, die nur Text enthält und kein herkömmlicher GIS-Datensatz, wie zum Beispiel eine Shape-Datei oder eine GeoJSON-Datei, wird auf Methoden der NLP zurückgegriffen. (Lampoltshammer & Heistracher 2012) Es ist jedoch anzumerken, dass textuelle Daten von GI-Systemen wenig beachtet wurden, da ein GIS vorwiegend zur Analyse von quantitativer Information herangezogen wird. Deshalb sind heutige GI-Systeme für eine textuelle Analyse nicht geeignet. Nichtsdestotrotz gibt es eine aktuelle Forschung, die sich dieser Thematik annimmt und versucht die Schnittmenge beider Forschungen herauszuarbeiten. (Paterson & Gregory 2019)

2.3 Kombination von GIS und Text

Ein Geoinformationssystem ist eine Art von Datenbank, die es erlaubt räumliche referenzierte Daten abzufragen, zu visualisieren und zu analysieren. Erreicht werden diese Funktionen dadurch, dass jede Zeile in der Tabelle für eine Position auf einer Karte steht. (Heywood, Cornelius & Carver 2011; DeMers 2008) Ein GIS braucht räumliche Daten – räumlich bedeutet, dass der Datensatz in einer Beziehung mit dem Standort in der realen Welt steht. Die Visualisierung erfolgt in Form einer Karte. Dabei wird eine Abstraktion der Realität vorgenommen. (Fazal 2008: 2)

Bei den Daten, die in einem GIS analysiert werden, handelt es sich um strukturierte Daten, während Texte keiner definierten Form folgen und unstrukturiert sind. Zur Kombination beider Datentypen gibt es eine nicht perfekte, aber ausreichende Lösung. Ein unstrukturierter Text wird in eine Tabelle aus Attributen konvertiert und mit räumlichen Informationen kombiniert. Damit diese beiden Forschungsfelder in der Praxis zusammen anwendbar sind, bedarf es zweier Voraussetzungen: Erstens muss ein Ortsname im Text herauszufinden sein und zweitens müssen Koordinaten zu diesem Punkt vorhanden sein. (Paterson & Gregory 2019) Sind diese Vorgaben in den Daten vorhanden so handelt es sich um *Geo-text Daten*. Als Beispiele zählen georeferenzierte Einträge von Facebook, Wikipedia, Zeitungsartikel, historischen Archiven oder andere Quellen, die eine Verbindung zwischen Ort und Text herstellen lassen. (Hu 2018) Geo-text Daten können mittels folgendem Ausdruck definiert werden: (Goodchild, Yuan & Cova 2007)

$$<f, [t], W> \quad (1)$$

In dieser Formel steht *f* für eine geographische Position. Es kann aber genauso gut für eine Linie (z.B.: einen Fluss oder eine Straße) oder auch für eine Fläche, wie einem Land, stehen. Außerdem kann es sich um vage definierte räumliche Objekte handeln. Dazu zählen kognitive Regionen oder andere Gebiete mit einer uneindeutigen Grenze. Die Zeit wird hier als optional angeführt, deshalb die Schreibweise *[t]*. In der Praxis werden dafür Zeitstempel, die eine Positionsänderung angeben, eingesetzt. *W* steht für den textuellen Inhalt, die Geo-text Daten beinhalten. Dazu zählen vorkommende Entitäten innerhalb des Textes, Themen, die ein Text beschreibt oder Sentiments. (Hu 2018)

Wenn eine räumliche Entität im Text identifiziert wurde, dann ist nur der Standort dieses einen Objektes bekannt. Weitere Analyse sind mit dieser Information nicht möglich. Aus

diesem Grund gilt es im nächsten Schritt auch die räumlichen Beziehungen zu extrahieren. In einem GIS werden drei unterschiedlichen Arten solcher Relationen unterschieden. Das sind die Topologie, die Distanz und die Richtung. (Egenhofer & Franzosa 1991) Die Beziehungen bestehen vorwiegend zwischen dem Punkt (ein 0-dimensionales Objekt), der Linie (bestehend aus mehreren Liniensegmenten, die mehrere Punkte in geordneter Reihenfolge verbinden) und der Fläche (Ein Bereich, der dessen Umgebung nicht miteinschließt). (Fegeas, Cascio & Lazar 1992) Diese geometrischen Objekte sind in Abbildung 10 visualisiert.

Die Topologie beschreibt die Beziehungen zwischen räumlichen Objekten. Dabei ändert sich deren Eigenschaften, trotz möglicher Transformationen nicht. (Kainz 2004) Um topologische Beziehungen zu klassifizieren hat sich das *9-Intersection Schema* durchgesetzt. Dabei werden zwei Geometrien (A, B) und deren drei Komponenten miteinander verglichen. Zu den Komponenten zählt der Rand (engl. boundary; b), das Innere (engl. interior; i) und das Äußere (engl. exterior; e). Daraus ergeben sich 3^2 unterschiedliche Kombinationsmöglichen, die in Abbildung 5 dargestellt sind. (Egenhofer, Mark & Herring 1994)

$$R(A,B) = \begin{pmatrix} A^b \cap B^b & A^b \cap B^i & A^b \cap B^e \\ A^i \cap B^b & A^i \cap B^i & A^i \cap B^e \\ A^e \cap B^b & A^e \cap B^i & A^e \cap B^e \end{pmatrix}$$

Abbildung 5 9-Intersection Matrix nach Egenhofer, Mark & Herring 1994

Die Distanz beschreibt den Aufwand der betrieben werden muss, um von einem Punkt zum anderen zu gelangen. Die Leistung kann entweder in Zeit, Längeneinheiten oder monetären Kosten angegeben werden. Eine andere Beziehung ist die Richtung. Diese ist eine orientierungsabhängige Beziehung zwischen einer Linie, welche zwei Objekte miteinander verbindet und einer fixierten Referenzlinie. Bei der Angabe von Richtungen wird meist auf kardinale Beziehungen zurückgegriffen. Darunter fallen die Angaben in Grad oder in Himmelsrichtungen. Je nach verfügbarer Granularität wird entweder auf ein 4-dimensionales (D_4 = {N, E, S, W}) oder 8-dimensionales (D_8 = {N, NE, E, SE, S, SW, W, NW}) System zurückgegriffen. (Sharma 1996: 33–36)

Im Bereich der natürlichen Sprachprozessierung werden üblicherweise die Topologie und die Richtung untersucht, da diese bei nutzergenerierten Daten am häufigsten vorkommen. Bei der Topologie handelt es sich um qualitative Relationsangaben, die als unscharf anzusehen und kompliziert auszuwerten sind. (Yuan 2011) Dem gegenüber stehen die Richtung und die Distanz, die sowohl qualitativ als auch quantitativ angegeben werden können. Dies hängt jeweils von der verwendeten Referenz ab. (s. Kapitel 2.4.4) Bezüglich der relevanten geometrischen Objekte in der NLP hat eine Untersuchung von Mark & Egenhofer 1994 gezeigt, dass Personen vorwiegend in Linien und Regionen denken. Das ist insofern nachzuvollziehen, als das Touristenhotspots oftmals eine Ausdehnung über mehrere Meter haben und bei der verbalen und schriftlichen Kommunikation nicht auf einen exakten Punkt fixiert wird. Für die praktische Aufarbeitung wird aber dennoch

festgelegt, dass die Distanzen aus den Beschreibungen und Bewertungen von Airbnb den Abstand zwischen zwei Punkten beschreiben.

In der natürlichen Sprache finden sich zwei Terminologien zu den räumlich relevanten Objekten, die eine räumliche Szene beschreiben. Das ist zum einen das *lokalisierte Objekt* oder die *Gestalt*, welches es zu lokalisieren gilt und zum anderen das *Referenzobjekt* oder der *Grund*, bei dem der Standort bekannt ist. Während die Gestalt sich bewegt bzw. theoretisch beweglich ist, erhält der Grund erst durch die Orientierung der Gestalt dessen Charakterisierung. (Talmy 1983) Ein Mehr an Eigenschaften ist in Tabelle 2 aufgelistet.

Assoziierte Eigenschaften	
Lokalisiertes Objekt	*Referenzobjekt*
Mehr beweglich	eher permanent verortet
Kleiner	Größer
Geometrisch einfacher	Geometrisch komplexer
Mehr von der Szene abhängig	Eher vertrauter und erwartet
Hohe Relevanz	Wenig Relevanz
Sticht eher hervor	Befindet sich im Hintergrund

Quelle: Talmy 1983

Tabelle 2 Eigenschaften von lokalisierten Objekten und Referenzobjekten

Der Ansatz von Talmy 1983 hat heutzutage immer noch Gültigkeit und wird vielfach angewandt, allerdings ist es nicht nachvollziehbar, dass eine Unterkunft, welche unbeweglich ist, sich laut dieser Definition bewegt. Ein neuerer Ansatz von Hu 2018 ist für diesen Fall logischer nachvollziehbar. Es handelt sich um die Überlegungen zu *Geo-text Daten*, die auf zwei verschiedene Arten entstehen können. Bei der sogenannten expliziten Methode schreibt eine Person über einen bestimmten Platz, der die derzeitige Position beschreibt. In der impliziten Art schreibt eine Person über einen Ort, an dem er oder sie sich zurzeit aber nicht befindet. Diese zwei Methoden sind aber nicht komplett voneinander abtrennbar, da es Überlappungsbereiche gibt. Je nachdem wie die Applikation trainiert ist, kann ein Ort über einen anderen gestellt oder gleichgestellt werden. (Hu 2018)

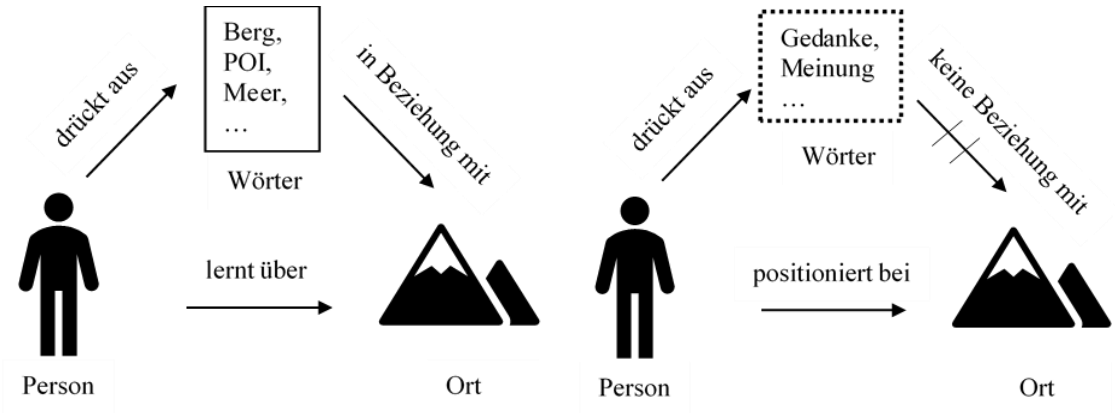


Abbildung 6 Erstellung von Geo-text Daten: explizit (l) und implizit (r) (Hu 2018: 5)

In der obigen Abbildung 6 sind die beiden Erstellungsarten visualisiert. Bei Bewertungen und Beschreibungen von Airbnb-Unterkünften handelt es sich um eine explizite Verortung. Die Personen befinden sich an dem Standort und sammelten Erfahrungen über diesen und verschriftlichten diese anschließend. Implizit würde in diesem Zusammenhang bedeuten, dass eine Person eine Unterkunft bewertet, an der sie sich nie aufgehalten hat.

2.4 Räumliche Information aus textuellen Daten

In dieser Thesis werden immer wieder Fachbegriffe verwendet, die für ein generelles Verständnis zu klären sind. Dazu zählen die Geosemantik, das Geoparsing oder die Fuzzy-Logik im Kontext von textuellen Daten. Außerdem sollen Datenfehler, die in diesem Zusammenhang auftreten können, anhand von aktueller Literatur dargelegt werden. Zuletzt wird festgestellt, wie es zurzeit möglich ist, einen geographisch relevanten Satz überhaupt zu identifizieren.

2.4.1 Geosemantik und Geoparsing

Bei der Semantik handelt es sich um die Deutung von Ausdrücken innerhalb einer Sprache. Sie stellt einen Kontrast zur Syntax dar. Zum Beispiel haben „Ich mag GIS“ und „Ich ♥ GIS“ eine eindeutig verschiedene Syntax, allerdings eine beinahe idente Semantik. Die räumliche Semantik oder auch Geosemantik fügt zwar die Restriktion ein, dass nur Ausdrücke mit einem Raumbezug von Interesse sind, allerdings ermöglicht es auch die Studien zur Untersuchung der Bedeutung von Standorten oder räumlichen Informationen. Nach Kuhn 2005 ist die Geosemantik als „*Verständnis von GIS-Inhalten und deren festhalten in formalen Theorien*“ zu definieren. Diese Definition kann in zwei Teile gegliedert werden: dem Verständnis und der Formulierung. Der erste Teil unterscheidet nach der Aufbereitungsart, die entweder für den Menschen (Verständnis von geographischen Konzepten und räumlichen Zusammenhängen) oder den Computer (Gazetteers oder semantische Interoperabilität zwischen verteilten Systemen) vollzogen werden kann. Ein Beispiel für die Formulierung wäre der Aufbau einer Ontologie, welche die Konzepte und Beziehungen ausspezifiziert. In einer Untersuchung ist es allerdings nicht notwendig, dass sowohl das Verständnis als auch die Formulierung erfüllt werden. (Hu 2017)

Geographisch relevant ist ein Satz oder Text nur dann, wenn er räumliche Information enthält. Dieser Schritt erfolgt meist über den Abgleich mit einem Gazetteer. Bei einem Treffer wird die Existenz von Ortsnamen bestätigt und dessen Eindeutigkeit festgestellt. (Hu 2017) Dieser Prozess wird als eine Art des Geoparsings beschrieben. Es hilft demnach die räumliche Sprache innerhalb von Texten zu erkennen und weist Wörtern einen Raumbezug zu, indem diese georeferenziert werden. (Moncla et al. 2014) Weiters sind auch alle Hinweise zur Positionsveränderung in Raum und Zeit, sowie zur Positionsverfeinerung als mögliches Ergebnis zu nennen. (Middleton et al. 2018) Es ist zu beachten, dass zwei verschiedene Ansichten zum Geoparsing existieren. Die eine

Strömung erachtet es als Teil eines Modells, um eine geographische Referenz zu bilden (Leidner & Lieberman 2011), während die andere Ausrichtung einen gesamten Prozess erklärt sieht, der auch eine automatisierte Geokodierung beinhaltet. Dieser Ansatz ist besonders bei Internetanwendungen von unterschiedlichen Anbietern, wie *Yahoo! Placemaker* oder *MetaCarta* im Einsatz. (Gritta et al. 2018; Gelernter & Balaji 2013) Für die vorliegende Arbeit ist demnach zweitgenannter Ansatz von höherer Bedeutung, da jeder einzelne Schritt abbildbar und praktisch umsetzbar ist. Die dafür benötigten Arbeitspakete sind in Abbildung 7 dargestellt: (Leidner & Lieberman 2011: 6–7)

- **Vorprozessierung:** Herausfiltern der textuellen Information, ohne dabei Angaben von Metadaten oder des Layouts zu berücksichtigen.
- **Geoparsing:** Finden von Ortsnamen, Beschreibungen oder jede andere Art von Information in einem Text, die eine geographische Relevanz aufweist.
- **Geokodierung:** Jedem gesonderten Ort wird eine Geokoordinate zugewiesen.
- **Räumliche Speicherung:** Für eine weitere Verarbeitung werden die Daten in einem Format gespeichert, welche eine geoinformationstechnische Weiterverarbeitung ermöglicht (z.B. Erzeugung eines räumlichen Index).
- **Räumliche Inferenz:** Prüfung ob eine räumliche Logik vorhanden ist. Das bedeutet, dass in diesem Schritt falsch geparste Information zu suchen sind.
- **Visualisierung:** Bildet den Abschluss einer jeden geographischen Applikation.

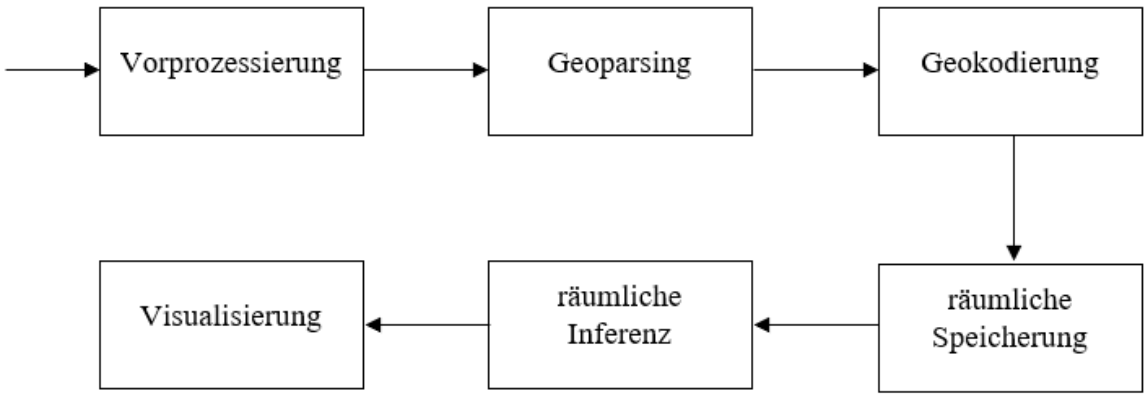


Abbildung 7 Prozess: geographische Relevanz aus Text (Leidner & Lieberman 2011: 6)

Im Kontext des Geoparsings werden Orte als Toponyme bezeichnet. Aus diesem Grund wird von nun an nur noch dieser Terminus verwendet. Zur Klärung sei gesagt, dass es sich laut den Vereinten Nationen bei einem Toponym um ‚*einen allgemeinen Namen für einen beliebigen Ort oder eine geographische Entität'* handelt. (Gritta et al. 2018)

Methoden des Geoparsings

Für die ‚namentlichen Objekterkennung' (engl. *Named Entity Recognition*, NER) gilt es, die „*Namen von Personen, Organisationen, geographische Position in einem Text zu*

erkennen“, des Weiteren auch „*zeitliche, monetäre und prozentuale Angaben*“ zu identifizieren. (Grishman & Sundheim 1996: 467) Demnach zählt auch das Geoparsing zu der übergeordneten NER.

Beim Geoparsing wird nicht nur eine Methode als vollkommen angesehen. In der Literatur werden drei unterschiedliche Ansätze der NER unterschieden und angewandt. (Gritta et al. 2018; Karagoz et al. 2016; Middleton & Krivcovs 2015) Die drei Ansätze werden nach datengetriebenen (Machine-Learning) und wissensbasierten Methoden (Gazetteer und Regeln; heuristischer Ansatz) differiert. (Moncla et al. 2017)

- Gazetteer-basiert:

  Bei diesem Ansatz wird ein geographisches Lexikon, ein sogenanntes Gazetteer angelegt. Die darin enthaltenen Toponyme werden mit den Token des abzugleichenden Satzes gegengeprüft. Ein Token beschreibt die einzelnen Wörter eines zerlegten Satzes. Den Inhalt des Gazetteers bestimmt der Anwendungsfall und der Ersteller selbst. In den meisten Fällen reichen Städtenamen und POIs. (Karagoz et al. 2016: 198) Im Internet gibt es auch schon vorgefertigte Gazetteers, wie *Geonames*. In dieser Datenbank sind zirka neun Millionen eindeutige Toponyme enthalten. Der Vorteil eines Gazetteers ist die simple Erstellung, trotz eines hohen Zeitaufwandes. Ein offenkundiger Nachteil ist, dass Toponyme, die nicht im Lexikon enthalten sind, auch nicht gefunden werden können. Zum Beispiel gibt es für die US-Stadt Washington, mindestens 64 verschiedene Referenzen. (Delozier, Baldridge & London 2015) Außerdem ist nicht gesagt, dass ein Treffer im Gazetteer, ein Wort im geographischen Kontext wiedergibt. Der Name „Chicago“ kann auf die US-amerikanische Stadt, die Pop-Band oder den internen Projektnamen von Windows 95 verweisen. Als letzter Nachteil sei gesagt, dass ein zu georeferenzierendes Objekt in unterschiedlichen Formen auftreten kann. Es wird sich keine Verbindung zwischen „EU“ und „Europäische Union“ herstellen lassen, wenn nicht beide in der Liste enthalten sind. (Purves et al. 2018)

- Regelbasiert:

  Ein Parser muss bei diesem ebenfalls heuristischen Ansatz ‚verstehen‘, welche zwei geographischen Typologien es gibt. Zum einen Informationen aus einem umfassenden Gazetteer, dass meist händisch angelegt ist und zum anderen ein Lexikon mit Konstrukten, wie zum Beispiel „nördlich von“, „entlang der“, „in“ oder „zwischen“. Somit wird in diesem Fall nach Mustern gesucht, um Toponyme zu finden. (Woodruff & Plaunt 1994) Vom Benutzer wird also festgelegt, welche Verben oder auch Pronomen auf ein Toponym folgen müssen. Das ist auch der wesentliche Nachteil, denn es kann zu falsch-positiven Erkennungen führen. Die Klassifizierung wird demnach zu empfindlich, denn bei der Phrase „zwischen den Tischen“, werden die Tische als zu georeferenzierende Objekte erkannt. (Karagoz et al. 2016: 199) Außerdem ist das Erstellen solcher Regeln aufwendig und immer wieder maßgeschneidert. Es ist anzumerken, dass ein regelbasierter-Ansatz zur namentlichen Objekterkennung auch für andere Fachbereiche, wie der Medizin oder der Finanz, zum Einsatz kommen kann. (Eftimov, Seljak & Korošec 2017)

- Machine-Learning-basiert:

  Um Toponyme zu erkennen wird bei Machine-Learning-Methoden, in diesem Kontext, meist ein überwachter Ansatz verwendet. In diesem Fall werden die bekannten Toponyme zuvor markiert. (Karagoz et al. 2016: 199) Es wird mittels unterschiedlicher Techniken, wie *Entscheidungsbäumen*, *Conditional Random Field* (CRF) oder *Support Vector Machines* (SVM) ein Modell auf Basis von selbsterstellten Trainingsdaten trainiert und mittels Testdaten validiert. Hierfür sind weder ein Gazetteer noch diverse Regeln von Bedarf, was als Vorteil anzusehen ist. Allerdings hängt das Ergebnis von der Güte der Trainingsdaten ab. Für diese Problemstellung sollte eine genaue Selektion an Beispielen erfolgen, die insgesamt balanciert und repräsentativ ist. Des Weiteren ist abzuklären, wie die Klassifikation von unbekannten Texten gehandhabt wird. (Purves et al. 2018; Khanam, Khudhus & Babu 2017)

Für alle Ansätze sei festgehalten, dass eine Normalisierung der textuellen Daten im Vorhinein notwendig ist. Meist besteht diese aus zwei Schritten: (1) Entfernen von sich unnötig wiederholenden Buchstaben, wie in „Straaaße“ oder „Schööööönbrunn“ oder weglassen von Buchstaben, wie bei „Stephdom“. Solchen Arten von Worten werden als „nicht-standardisierte“ Wörter beschrieben. Bei Abkürzungen kann meist davon ausgegangen werden, dass es sich um absichtliches Weglassen von Buchstaben handelt. (2) Außerdem gilt es morphologische Eigenheiten der analysierten Sprachen zu berücksichtigen. Beispiele hierfür sind Umlaute oder Abkürzungen. Besonders bei der Auswertung von Texten mit Umlauten ist auf eine korrekte Kodierung zu achten. Als Ziel haben alle Anwendungen, dass eine Standardisierung des Textes passiert. (Karagoz et al. 2016: 199)

Wenn ein Gazetteer als Nachschlagewerk bei der NER verwendet wird, dann sind Toponyme nur durch einen Treffer beim Abgleich zu identifizieren. Bei diesem Vorgehen sind drei wesentlich Problemstellungen zu beachten, die jeweils mit einem heuristischen Ansatz zu lösen sind: (Nadeau 2007: 42–44)

- Entität-Nomen Unsicherheit

  Diese Uneindeutigkeit tritt dann auf, wenn die Entität das Homonym des Nomens ist. In Nadeau 2007 wird eine Heuristik von Mikheev, Moens & Grover 1999 erläutert, die allerdings vorwiegend in der englischen Sprache eine Anwendung findet. Im Englischen werden Wörter am Anfang eines Satzes und wenn es Eigennamen sind (z.B.: geographische Namen, Sehenswürdigkeiten oder persönliche Namen), großgeschrieben. Neben dem ersten Wort eines Satzes kann es jedoch zu Unsicherheiten kommen, denn das kleingeschriebene „jobs“ steht für eine Arbeitsstelle, während „Jobs“ ein Nachname ist. In der deutschen Sprache ist eher von einer Entität-Verb Unsicherheit zu sprechen. In Wien gibt es die Adresse „[Am] Graben“, aber auch die handwerkliche Tätigkeit „graben“.

- Entitätsgrenzen-Erkennung

  Bei der Verwendung eines Nachschlagewerks gibt es üblicherweise bei Entitäten das Problem, dass es unsicher ist, wo es anfängt und wieder endet. Das kann dann passieren, wenn die Entität aus mehr als einem Wort besteht. Zum Beispiel ist „Paris" auch in „Paris Saint Germain" enthalten – dies verweist nicht auf eine Stadt, sondern auf einen Sportverein. Des Weiteren sind Unsicherheiten möglich, wenn die umgebenden Worte unbekannt und großgeschrieben sind (vor allem im Englischen). Bei der Zeitung „Washington Post" folgt der Satz „Democracy Dies in Darkness", welcher lt. der Orthografie unnötigerweise großgeschrieben ist. Nach Palmer & Day 1997 ist zu empfehlen, dass alle aufeinanderfolgenden Wörter, welche den gleichen Typ haben (z.B.: Subjekte) oder Entitäten, die von großgeschriebenen Wörtern umgeben sind, miteinander verbunden werden.

- Entität-Entität Unsicherheit

  Tritt immer dann auf, wenn eine Entität zu mehr als einem Typ passt. Das Wort „Memphis" ist sowohl ein Vorname als auch eine Stadt in den USA. Petasis et al. 2001 empfiehlt, dass in diesem Fall das lokale Dokument als Gesamtes untersucht werden soll. Wenn innerhalb des Dokuments an einer anderen Stelle der Kontext von „Memphis" hervorgeht, wie zum Beispiel bei „Memphis Depay", dann kann davon ausgegangen werden, dass es sich um eine Person, in diesem Fall einen Fußballspieler handelt. Daraus kann abgeleitet werden, dass der alle vorkommenden „Memphis" im Text sich nicht auf die Stadt beziehen – eine klarere Typzuordnung ist nun möglich.

Geokodierung

Bei der Geokodierung handelt es sich um einen Prozess, der auf Basis einer textuellen postalischen Adresse, diese in eine geographische Koordinate umwandelt. Nach erfolgter Geokodierung kommt der Vorteil von einem GIS zum Tragen, da sobald eine räumliche Referenz vorhanden ist, auch eine Verbindung mit anderen Geodaten aus anderen Quellen möglich ist. (Barnes 1995: 908)

Dieser Weg zu einer geographischen Koordinate enthält allerdings zwei wesentliche Problemstellungen. (1) Die Adressen können auf verschiedene Arten veröffentlicht werden. Menschen schreiben die Anschriften unterschiedlich und damit in einer nicht-standardisierten Form. Um trotzdem zu einem Ergebnis zu gelangen werden komplizierte Algorithmen, als einfacher Textabgleich benötigt. (2) Eine Möglichkeit das obige Problem zu beheben ist die Überlegung zur Standardisierung von Adressbestandteilen mittels eines Wörterbuchs. Somit könnten alle unterschiedlichen Schreibweisen von Straße, wie „Strasse", „Str." und „St." auf dieses Wort referenzieren. Dadurch entsteht jedoch das Problem, dass alle Notationen bekannt sein müssen, ansonsten kann auch hier keine Zuordnung erfolgen. (Xu, Flexner & Carvalho 2012)

2.4.2 Fuzzy-Logik beim Geoparsing

Wie im vorherigen Kapitel 2.4.1 erläutert, kann es im Rahmen von Abgleichen mit geographisch referenzierten Standorten zu Unschärfen (engl. *Fuzziness*) kommen. Die Toponyme können entweder falsch erkannt oder falsch geschrieben werden. Die beiden Fehlerquellen sind zu einem gewissen Grad auch voneinander geprägt. Auf Basis einer Zugehörigkeitsfunktion wird bestimmt, zu welchem Fuzzy-Set ein Element gehört. Der Zugehörigkeitswert liegt zwischen Null und Eins. (Kainz 2007) Die Definition, welcher Satz für die räumliche Analyse von Relevanz ist oder nicht, wird in Kapitel 3.2.4 festgelegt. Für diese Arbeit wird keine eigene Zugehörigkeitsfunktion im Sinne der Fuzzy-Logik erstellt, allerdings wird mit diesem Kapitel auf dessen theoretisch mögliche Verwendung im Geoparsing-Prozess hingewiesen.

Die Notwendigkeit zur Klärung von Fuzzy-Logik in einem solchen Zusammenhang, ist dadurch gegeben, dass Thurgar-Dawson 2014 drei unterschiedliche Schwierigkeitsstufen zur Findung von Toponymen unterscheiden. In der mittleren, der unsicheren Stufe B, soll mit Hilfe des Prinzips der Fuzzy-Logik eine Lösung gefunden werden.

- Stufe A

  In diesem Set befinden sich all jene Wörter, die eine hohe geographische Relevanz und obendrein noch einfach zu lokalisieren sind. Innerhalb des Satzes heben sie sich eindeutig vom Rest der Objekte hervor.

  Beispiel: *Bei einem Besuch in **Wien** buchen sie immer ein Hotel in der Nähe des Stephansdoms.*

- Stufe B

  Das Georeferenzieren auf einer Karte ist bei diesen Worten schwierig, aber durchaus möglich. Es ist unsicher auf welchen Standort sich der Text bezieht. Eine erfolgreiche Geokodierung wird durch eine Mischung aus Fuzzy-Logik und geosprachlicher Intelligenz erreicht.

  Beispiel: *Am Flughafen **Berlin Tegel** erwartete ich die Ankunft ihres Fliegers aus **New York**.*

- Stufe C

  In einem solchen Fall gibt es keine Möglichkeit eine Georeferenzierung sinnvoll umzusetzen, da ein Finden von Toponymen nicht möglich ist. Ein Satz kann in einem solchen Fall überhaupt keine oder zu viele Ortsbezeichnungen enthalten.

  Beispiel: *Ich war mir nicht mehr sicher wie die Stadt hieß, in der ich mich befand: entweder **Linz**, **Graz** oder doch **Innsbruck**.*

In der menschlichen Sprache werden räumliche Fragen meist mit qualitativen Aussagen, „nah“, „fern“, „weit“, etc., beantwortet, da diese meist intuitiver sind. Das qualitativ-räumliche Denken von Menschen basiert meist auf einer egozentrischen Karte in deren

Gedanken. Von einem zentralen Punkt aus werden Beziehungen zu umgebenden Referenzobjekten gezogen. (Gahegan 1995) Diese Angaben sind allerdings in keinem GIS umsetzbar, da diese wie weiter oben beschrieben (Kapitel 2.3) auf quantitative Daten zurückgreifen. Es herrscht eine Unschärfe in den Daten, die nicht ohne weiters interpretierbar ist. Außerdem können keine geometrischen Entfernungen angenommen werden, denn das Verständnis von qualitativen Entfernungen ist zu unterschiedlich. Des Weiteren gibt es Ortsnamen, die sehr vage sind, da deren räumliche Ausdehnung nicht mit der realen Grenze übereinstimmt. Ein mögliches Beispiel ist die Angabe der Entfernung zur Innenstadt. Es ist unsicher, ob die Person bei einer schriftlichen Beschreibung das Stadtzentrum, den Zentroid oder die äußere Grenze der Innenstadt bereits als diese ansieht. (Purves et al. 2018) Weiterführende Information dazu sind aus Wallgrün, Klippel & Baldwin 2014; Spranger & Pauw 2012 zu entnehmen.

2.4.3 Schwierigkeiten mit nutzergenerierten Daten

Textuelle Daten, die von Menschen generiert wurden, wie Blogs, Tweets oder Rezessionen, halten sich meist nicht an strenge Regeln. Beispiele sind die sprachliche Diversität, die Mehrdeutigkeit von Ortsnamen, eine metonymische Sprache oder das fehlende Verständnis von Kontext. (Gritta et al. 2018) Bis auf den letzten Punkt sind alle Punkte mit einem hohen menschlichen Einsatz lösbar. Was in der Aufzählung von Gritta et al. 2018 allerdings fehlt, ist auch die Möglichkeit von falsch erkannten oder geschriebenen Wörtern.

Für die spätere Methodenentwicklung gilt es zu identifizieren mit welchen sprachlichen Fehlerarten der Prozess konfrontiert sein wird. Auf Basis dieser Überlegungen ist es essenziell Lösungsstrategien zu entwickeln. Bei der Studie von Middleton et al. 2018 wurden Twitterdaten ausgewertet, die mit dem Aufkommen von Notsituationen (Terrorangriff oder Naturkatastrophe) korreliert sind. Mit einer Vielzahl an Posts wurden Fehlermuster festgelegt, sowie der Auftretungswahrscheinlichkeit angegeben. In Tabelle 3 wird nur auf die häufigsten Fehler eingegangen.

Fehlermuster	Auftreten	Echt-Beispiel	Korrekte Darstellung
Übliche Begriffe als Toponym	Sehr üblich	This is the end of my Hurricane Sandy live tweeting day 1	Keine. Hurricane wurde als Stadt in den USA verortet
Personennamen, die auch Standortnamen sind	Üblich	Webgrrls hosting company is flooded by #Sandy	Sandy, UT, USA
Standort ohne Kontext	Üblich	The city has high winds and flooding by the coastal lines	Innenstadt von London, UK
Keine gut formatierte Adresse	Selten	Street flooding #NYC 48th Ave	48th St, New York, NY, USA
Rechtschreibfehler	Selten	Earthquake in Chrristchurch New Zealand ghastly	Christcurch, Neuseeland

Quelle: Middleton et al. 2018

Tabelle 3 Fehlermuster von Tweets mit Standortbezug

Es gilt in der Auswertung der Ergebnisse zu überprüfen inwieweit die angeführten Fehlermuster für diese Arbeit von Relevanz sind. Wie schon in Kapitel 1.3 erwähnt, werden Geoparsing-Prozesse im Mikroblog-Bereich meist bei Katastrophensituationen eingesetzt. Diese Texte werden unter mentalen Stress erstellt und nicht zwingend einer Korrekturlesung unterzogen. Ähnliche Untersuchungen sind neben der von Middleton et al. 2018, auch in Lingad, Karimi & Yin 2013 und Gelernter & Mushegian 2011 enthalten. Zu Fehlermuster in textuellen Daten, die in einer vermeintlich angenehmeren und ruhigeren Situation verfasst worden sind, gibt es keine Literatur, welche die Unstimmigkeiten so feingranular wiedergibt.

Die Textnormalisierung, welche in Kapitel 2.4.1 erläutert wurde, ist ein Teil einer Datenbereinigung, denn ohne dieser wäre die Anwendung von NLP-Techniken nicht möglich. Je nach Quelle des Textes bedarf es einem entweder kleinen oder großen Normalisierungsaufwand. Ein Vergleich von Einträgen aus dem sozialen Medium „Twitter“ mit Zeitungsartikeln, Blog- oder Forumseinträgen, lässt Unterschiede in der Erstellung, als auch in der Wortlänge, sowie der grammatikalischen Richtigkeit feststellen. Bei Twittereinträgen ist die Zeichenlänge auf 280 beschränkt und es kann keine Bearbeitung nach Absetzung der Nachricht erfolgen. Im Gegensatz dazu ist eine nachträgliche Editierung von Blogeinträgen oder Wikipedia-Artikeln möglich. Im Falle von Wikipedia und Zeitungsartikeln wird meist auch eine Kontrolle durch einen oder mehrere Bearbeiter durchgeführt. (Baldwin et al. 2013)

Die verschiedenen Eigenheiten der angegebenen Korpusse sind in der Tabelle 4 und der Tabelle 5 dargelegt. Es zeigt sich, dass wenn keine Zeichenbeschränkung vorliegt die Wort-, als auch die Satzlänge zunimmt. Außerdem dürften Nutzer auf Wörter verzichten, die nicht in allgemein gültigen Lexika zu finden sind. (engl. *out of vocabulary,* OOV) Des Weiteren ist für das Parsen eine grammatikalische Richtigkeit von Bedarf. Ist dies nicht

gegeben, so kann auch kein Rückschluss auf die Semantik geschlossen werden. Es zeigte sich, dass die längeren Sätze, die in Wikipedia oder in der Zeitung eine Anwendung finden auch schwieriger für den Computer sind zu parsen. Nichtsdestotrotz ist der prozentuale Anteil an nicht-parsing-fähigen Sätze geringer – allerdings ist der Unterschied zu den nicht editierten Texten geringer als angenommen. (Baldwin et al. 2013)

Korpus	**Wortlänge**	**Satzlänge**	**%-OOV**
Twitter	3,8 ± 2,4	9,2 ± 6,4	~ 23
Forum	3,8 ± 2,3	14,2 ± 12,7	~ 17,5
Blog	4,1 ± 2,8	18,5 ± 24,8	~ 20
Wikipedia	4,5 ± 2,8	21,9 ± 16,2	~ 19
Zeitung	4,3 ± 2,8	19,8 ± 14,5	~ 16,8
Quelle: (Baldwin et al. 2013)			

Tabelle 4 Allgemeine morphologische Gegebenheiten von Texten verschiedener Korpusse

Korpus	**kein parsing möglich [%]**
Twitter	37,4
Forum	35,6
Blog	35,3
Wikipedia	26,2
Zeitung	23,2
Quelle: (Baldwin et al. 2013)	

Tabelle 5 Anteil an Sätzen, die durch undefinierbare Grammatik nicht *parsebar* sind

Mit welchen der obigen Quellen, die Angaben in Airbnb am meisten Ähnlichkeit aufweisen, wird sich weisen. Die von den Gästen und Hosts erstellten Texte haben in diesem Fall keinen wissenschaftlichen Anspruch und können nach der Veröffentlichung auch nicht mehr geändert werden. Eine vorherige Überarbeitung durch eine zweite oder dritte Person kann nicht nachgewiesen werden. Nach diesen Annahmen dürften Beschreibungen und Bewertungen von Airbnb auf dem morphologischen Niveau von Foren, Blogs oder Wikipedia-Einträgen rangieren.

2.4.4 Räumliche Beziehungen in der Sprache

In den Sprachwissenschaften hat sich der Terminus der ‚räumlichen Sprache' entwickelt. Diese hat immer das Ziel ein Objekt zu verorten. Die sprachliche Repräsentation eines Ortes benötigt drei Elemente: (1) ein Referenzobjekt, (2) das zu verortende Objekt und (3) die dazugehörige räumliche Beziehung, welche meist als Präposition oder Distanz angegeben wird. (Vasardani et al. 2013) Meist werden mittels Präpositionen räumliche Beziehungen hergestellt. In einem solchen Fall dienen diese als topologische Einschränkung. (Moncla 2016: 17) Wenn die drei obigen Elemente in einem Satz vorkommen, so wird von einer räumlichen Dreiergruppe (engl. Triplet) gesprochen. Dieser Ansatz der Attribuierung hat sich als Standard in der räumlichen Semantik durchgesetzt. Beim Kurznachrichtendienst Twitter sind zirka 2% aller Token einem räumlichen

Ausdruck zuzuordnen. (Liu, Vasardani & Baldwin 2014) Ein *Triplet* ist die Idealform der räumlichen Ausdrucksweise. Wenn der Korpus allerdings aus informellen Quellen stammt, wie aus einem Dialog oder sozialen Medium, dann sorgen meist die *verkümmerten lokalen Ausdrücke* für eine niedrigere Anzahl an relevanten Sätzen. In einem solchen Fall fehlt das Subjekt, auf das sich die Präposition bezieht. Als Beispiel kann der Satz „bei der Universität" angewandt werden. Die Frage nach dem Wer oder Was ist bei der Universität kann nicht eindeutig beantwortet werden. (Khan, Vasardani & Winter 2013) Die Auflösung von verkümmerten Ausdrücken steht nicht im Zentrum dieser Arbeit, deshalb werden solche Sätze als „nicht relevant" gekennzeichnet.

Dieser räumliche Zusammenhang existiert in der realen Welt nicht, sondern hilft dem Menschen die Welt zu verstehen und eine Interaktion zu ermöglichen. Es werden zwei Ansätze unterschieden, um ein räumliches Verhältnis abzuleiten. Entweder es basiert auf Erfahrungen oder auf formalen Gesetzmäßigkeiten, wie aus der Mathematik oder Logik. Ersteres Modell ist leichter zu erlernen, da Individuen deren Umwelt meist auf die gleiche oder ähnliche Art erkunden. Das formale Modell ist schwieriger, da nicht jede Person zum Beispiel die mathematische Grundlage für die Berechnung der euklidischen Distanz besitzt. (Mark & Frank 1989: 539)

Weiters können räumliche Beziehungen dazu beitragen, dass auf Basis einer lokalisierten Entität die bisher unbekannte Position einer anderen Entität gefunden wird. Um die Position eines Objektes innerhalb einer Beziehung auszudrücken werden drei verschiedene Typen eines ‚Bezugssystems' unterschieden: intrinsisch, absolut und relativ. (Moncla 2016: 17)

- Intrinsische Beziehung: Die Beschreibung des Referenzobjektes bezieht sich auf dessen inhärenten Eigenschaften. Im angegebenen Beispiel ist es irrelevant wo sich der Betrachter befindet, da das Vorne und das Hinten der Kirche festgelegt ist.

  Beispiel: *Die Statue befindet sich vor der Kirche.*

- Absolute Beziehung: Hier werden kardinale Richtungen angegeben, um die relative Position einer Entität zu finden. Diese räumliche Beziehung ist von einer unabhängigen, externen Referenz abhängig.

  Beispiel: *Die Hütte befindet sich nördlich des Sees.*

- Relative Beziehung: In diesem Fall ist es wichtig zu wissen, wo sich der Beobachter zum Referenzobjekt befindet. Die Interpretation des Textes hängt von dessen Orientierung ab.

  Beispiel: *Die Hütte befindet sich links des Sees.*

Die obigen Begrifflichkeiten beschreiben räumliche Beziehungen zwischen punkt-, linien- und flächenhafte Darstellungen/Objekten im Allgemeinen. Eine weitere Beziehung von räumlichen Objekten stellt die Topologie dar. Die Topologie beschreibt vorwiegend die Beziehungen von räumlichen Regionen zueinander. Für diese Arbeit von Interesse ist die partielle Überlappung von Flächen. Wie aus dem Namen hervorgeht, überschneiden sich

hier die Flächen zweier oder mehrerer Standorte. Zum Schluss dieses Kapitel sei noch auf die metrische Beziehung hingewiesen. Die Distanzen folgen einer euklidischen Repräsentation und die daraus entstehende quantitative Darstellung sorgt für eine genauere Messung zwischen Geoobjekten („*Fahren Sie die nächsten 400m geradeaus"*). (Moncla 2016: 18–20)

2.5 Menschliche Wahrnehmung von Distanzen

Je nach Verwendung von unterschiedlichen Präpositionen in einem Satz ergeben sich verschiedene Arten von räumlichen Beziehungen. Die in Tabelle 6 aufgelisteten Familien von räumlichen Beziehungen und deren Beispielwörter helfen einen nicht bekannten Standort auf Basis eines bereits georeferenzierten Standortes zu finden.

Familien von räumlichen Beziehungen	**räumliche Beziehung**
kardinale Richtungen	Nord, Süd, Ost, West, Nordost, Südwest, Südost, …
qualitative Distanzen	nah, fern
relative Richtungen	davor, dahinter, links davon, rechts davon
topologisch	innerhalb, überlappend, überdeckt, gleich, nicht verbunden
Quelle: (Chen, Winter & Vasardani 2018)	

Tabelle 6 Familien von räumlichen Beziehungen

Die spätere praktische Auswertung basiert auf nutzergenerierten Daten. Jeder von diesen Personen hat eine eigene räumliche Wahrnehmung. Für manche Menschen sind 500m eine große Distanz, für andere wiederum sind es erst 1.000m. (s. qualitative Distanzen) Obwohl in dieser Arbeit ausschließlich quantitative Angaben ausgewertet werden, soll dieses Kapitel darlegen worauf die menschliche Wahrnehmung von Distanzen beruht und welche Schwierigkeiten bei der späteren Analyse zu erwarten sind. Es könnten nicht nur unscharfe qualitative Angaben, sondern auch fehlerhafte bzw. widersprüchliche quantitative Aussagen geschrieben worden sein.

Für die Analysen von textuellen Informationen mit Raumbezug wird, meist unbewusst, ein Suchraum festgelegt. Dessen Begrenzung kann aber auch in Form eines *kontextuellen probabilistischen Suchraum*-Modells definiert sein. Ein solches Modell hilft beispielweise bei der Erstellung eines Gazetteers, da festgelegt wird, auf welcher Granularität die Einträge basieren. Für die Erstellung sind zwei Begriffe wesentlich: das Relatum und das Lokatum. Das Relatum wird auf Basis des Lokatum georeferenziert. Die Position des Lokatum ist also bereits bekannt. Mit vorhandenen Trainingsdaten und der Berücksichtigung von vier Faktoren lässt sich ein solches Suchraum-Modell erstellen: (Chen, Winter & Vasardani 2018)

- Granularität des Relatum

  Das Relatum befindet sich mit dem Lokatum und der Präposition in einer Dreiergruppe. Je nachdem, ob „nahe dem Restaurant“, „nahe dem Flughafen“ oder „nahe von Wien“ verwendet wird, ist ein anderer Suchraum bei der Interpretation festzulegen. Es werden fünf verschiedene Granularitätsstufen unterschieden: Gebäude, Straße, Bezirk, Stadt und über die Stadt hinaus.

- Granularität des Lokatum

  Dieser Faktor entspringt dem gleichen Ansatz wie vorher. Je nachdem, ob das Lokatum ein Gebäude oder ein ganzer Vorort ist, wird ein anderer Suchraum definiert.

- Bekanntheit des Relatum

  Je prominenter das Relatum ist, desto eindeutiger ist dessen Position. Je öfter ein Ort in den Trainingsdaten vorkommt, desto bekannter ist dieser. Ein prominentes Relatum hilft meist, weniger bekannte Standorte zu finden.

- Granularität des räumlichen Diskurses

  Je nach Diskurs können Ortsbeschreibungen verschieden interpretiert werden. Es ist zu empfehlen, dass nur Informationen analysiert werden, die dasselbe Level an Granularität aufweisen. So haben zum Beispiel das Relatum und das Lokatum im folgenden Satz dieselbe Granularität: „Das Hotel liegt hinter dem Stephansdom“ (Level: Gebäude-Niveau).

Diese Problemstellung von wahrgenommen Distanzen wird als kognitive Entfernung beschrieben. Diese erzeugt Schätzungen (gemessen in Zeit, Länge oder Kosten) zu Orten, die zurzeit der Beschreibung nicht sichtbar sind. Die Untersuchung von kognitiven Distanzen sieht ihre Relevanz darin begründet, dass diese einen hohen Einfluss auf das Reiseverhalten hat. Es wird analysiert, (1) ob überhaupt gegangen oder stehengeblieben werden soll, (2) wohin gegangen werden soll oder (3) welche Route genommen werden soll. (Walmsley & Jenkins 1992a) In dieser Arbeit wird nur auf die Wahrnehmung von Touristen eingegangen. Als Touristen werden Personen bezeichnet, die einen Ort nur temporär besuchen und der weniger vertraut ist, als die heimische Umgebung. (Walmsley & Jenkins 1992b) Eine unbekannte Stadt ist für Menschen zu schwierig einzuprägen, aus diesem Grund erfolgt eine innere, subjektive Vereinfachung des Gesehenen. Die Informationen werden für einen späteren Wiederaufruf minimalst „gespeichert“. Das Lernen von Orten beinhalten vier Arten des Wissens: (Briggs 1973)

- Kenntnis von Orten (z.B. Einkaufsmöglichkeiten)
- Näheverhältnis zwischen dem Individuum und Orten (Angabe als Distanz)
- Verortung von Orten in Relation zueinander (Angaben zur Richtung)
- Wissen über eine Menge an Orten und deren Distanz zueinander

Dieses Bild, das dabei im Gehirn eines Menschen entsteht, wird „Mental Map“ genannt. Eine beispielhafte Mental Map eines Touristen der zwei Wochen in einer für ihn fremden Stadt verbrachte ist in Abbildung 8 dargestellt. Eine Mental Map besteht im Wesentlichen aus fünf Elementen. Dazu zählen Pfade, Kanten, Bezirke, Knoten und Sehenswürdigkeiten. Die Pfade beschreiben jene Strecken, auf denen sich die Personen im Raum bewegen. Als Kanten werden vorwiegend die Grenzen empfunden, die Bereiche und Nachbarschaften voneinander trennen. Die Bezirke zeichnen sich dadurch aus, dass sich diese in den Punkten der Charakteristik, der Geschichte und anderer Eigenschaften ähneln. Ein Knoten beschreibt ein Zusammentreffen von mehreren Pfaden. Als Beispiele sind Parks oder große Plätze zu nennen. Die Sehenswürdigkeiten selbst repräsentieren eine Stadt, denn sie haben meist ein Alleinstellungsmerkmal im weltweiten Vergleich. Je nach Anwendergruppe sind unterschiedliche Elemente einer Mental Map stärker ausgeprägt. Für die Touristen sind die Sehenswürdigkeiten einprägender, während für die ansässige Bevölkerung, die einzelnen Pfade einen höheren Wert für die Orientierung leisten. (Lynch 1960) Zum Extrahieren bzw. Verschriftlichen der Information werden unterschiedlich granulare Techniken angewandt. Vom Zeichnen einzelner Sehenswürdigkeiten, bis hin zur Beschreibung der kardinalen Richtungen inklusive Distanzabschätzungen. (Walmsley & Jenkins 1992b) Es hat sich gezeigt, dass die Einschätzungen von Fußgänger anhand der Mental Map näher an die Realität kommt, als jene von Personen, die nur mit dem motorisierten oder öffentlichen Verkehr unterwegs sind. (Young 1999) Als weitere wichtige Erkenntnis kann mitgenommen werden, dass Personen, die mit ihrer Umgebung vertraut sind, eine höhere geometrische Genauigkeit bei den kognitiven Distanzen aufweisen. Für die Analyse von Angaben von Touristen gilt also: Je länger der Aufenthalt, desto präziser sind die Angaben aus der Mental Map. (Golledge & Stimson 1998)

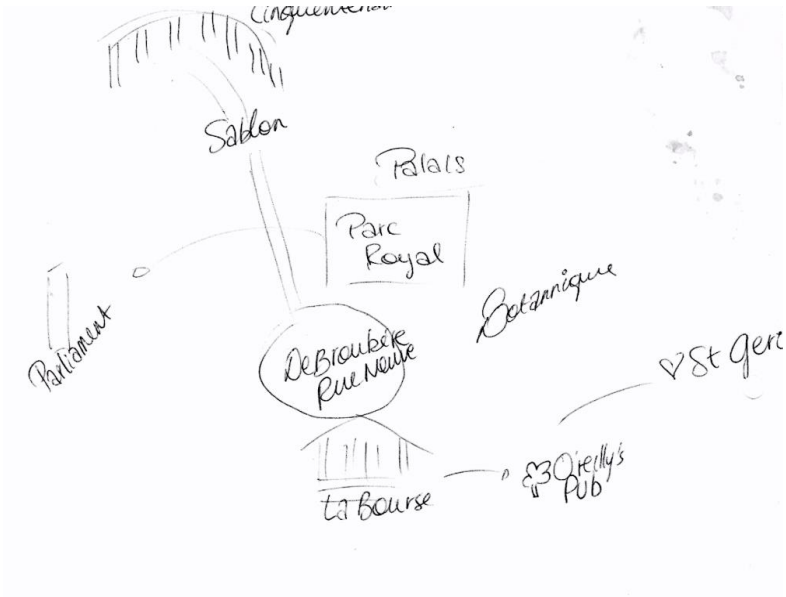


Abbildung 8 Eine Mental Map eines Touristen aus Brüssel (Hofmann 2012: 60)

Da es sich um nutzergenierte Daten von Touristen handelt sind noch Annahmen zur Interpretation der Distanzangaben zu tätigen. Die folgenden beiden Absätze widmen sich

deshalb der Fortbewegungsmethode, der Geschwindigkeit, als auch der kognitiven Messung der Entfernung.

Die Stadt Wien verfügt über das sechstgrößte Schienenbahnnetz der Welt und ist generell kleinräumig strukturiert. Alle Sehenswürdigkeiten sind mit den öffentlichen Verkehrsmitteln erreichbar. Aus diesem Grund wird angenommen, dass Touristen, so wie die ansässige Bevölkerung vorwiegend mit dem öffentlichen Verkehr unterwegs sind. Diese Annahme stützt sich auf einer Studie, die aufklärt, dass die Wahrscheinlichkeit bei einem innerstädtischen Ortswechsel in Wien ein Auto zu verwenden, von 37% (2010) auf 27% (2015) gesunken ist. (Stadt Wien 2016) Sofern nichts anderes angegeben ist, beziehen sich die Angaben der Touristen und Hosts auf Airbnb auf die fußläufige Distanz. Beispielhafte Angaben wären hierzu Wörter, die eindeutige darauf verweisen, dass mit der U-Bahn, dem Auto oder dem Rad gefahren wurde.

Abschließend gilt es zu bestimmen wie Distanzen von Menschen angegeben werden. Die Angabe kann sowohl als Isochrone, also auch mittels Luftlinie erfolgen. Als Isochrone werden Linien angegeben, die zeigen wie weit in einer bestimmten Zeit gegangen werden kann. Dabei ist das zugrundliegende Verkehrsnetzwerk oftmals unterschiedlich detailliert ausgearbeitet. Es können Ampeln, Steigung, Streckenlänge und Fußgängeraufkommen einen Einfluss auf die zurücklegbare Distanz haben. Die Modellierung eines hochgranularen Netzwerkes steht im Rahmen dieser Arbeit nicht im Vordergrund, deshalb wird angenommen, dass es keine Hindernisse, wie Ampeln, Steigungen oder weitere Fußgänger gibt. Eine weitere Vereinfachung der Modellierung stellt die „Umwegempfindlichkeit“ von Fußgängern dar. Für das später angewandte Verkehrsnetzwerk wird aus diesen Gründen pro Distanzangabe eine Isochrone gebildet, die auf der kürzesten Strecke bei konstantem Tempo basiert. Die Geschwindigkeit von Fußgängern ist in diesem Zusammenhang relativ eindeutig. Diese beträgt zwischen vier und sechs Kilometer pro Stunde, also im Durchschnitt fünf km/h. (Frey 2015) In Abbildung 9 ist der Unterschied zwischen Isochronen und Luftlinien dargestellt. Für die Erstellung von Luftlinien wird ein weniger komplexer Algorithmus benötigt, als bei Isochronen. Bei der Erstellung von Luftlinien wird angenommen, dass es keine Hindernisse gibt und eine Ausbreitung in alle Richtungen möglich sind (Isotropie). (Weibel 2016)

Abbildung 9 Isochrone (l) im Vergleich mit Luftliniendistanzen (r) (Wilke 2013)

2.6 Erreichbarkeitsanalysen

Wenn die Geo-text Daten aus dem Korpus ermittelt sind und daraus dann die räumlich relevante Information, in Form von Entfernungsangaben extrahiert ist, so gilt es diese Zahlen in einem GIS zu visualisieren und zu analysieren.

Im Rahmen einer solchen Analyse werden alle drei Datentypen eines Vektors angewandt. Das ist der Punkt, der den Point of Interest (POI) wiedergibt, die Linie, welche die einzelnen gewichteten Straßenelement des Verkehrsgraphen darstellt und zuletzt die Fläche, als Einzugsgebiet rund um den POI. (Abbildung 10) Es ist dabei unerheblich, ob es sich um Luftlinien oder Isochrone handelt. Das Resultat eines solchen Vektorpuffers ist immer eine Fläche bzw. ein Einzugsgebiet. Wenn sich mehrerer solcher Puffer überlappen, so wird von geschachtelten Distanzzonen gesprochen. (Weibel 2016)

Abbildung 10 Vektordaten: Punkt (a), Linie (b) und Fläche (c) (Paterson & Gregory 2019: 45)

Ein Einzugsgebiet ist Teil einer Erreichbarkeitsanalyse. Dieser Forschungsgegenstand kann in vier Unterkategorien geteilt werden: (Hass 2009: 7ff)

- Einzugsgebiets-Messungen
  - Container: Zählung bestimmter Einheiten in einem Gebiet
  - kumulative Möglichkeiten: Zählung von Möglichkeiten innerhalb einer bestimmten Zeit oder Distanz
- Reiseimpedanz-Messungen
  - Minimale Distanz: Entfernung zur nächsten Einrichtung
  - Reisekosten: Die durchschnittliche Distanz zwischen jedem Quell- und Zielpunkt
- Anziehungs- und Potentialmessungen
  - Potential: Die Summe aller Einrichtungen (nach Größe gewichtet) wird durch einen eigens festgelegten Effekt der Distanz dividiert
- Messungen basierend auf der Nützlichkeit
  - Nützlichkeit: Die Wahrscheinlichkeit, dass sich eine Person für eine bestimmte Einrichtung entscheidet, hängt von subjektiven Standpunkten ab

Auch wenn für die Arbeit, aufgrund der Anforderungen, nur die Reiseimpedanz-Messungen in Frage kommen, so erfolgt die Darstellung der Ergebnisse in einem Einzugsgebiet. Die Erstellung eines Einzugsgebiets zählt im Kontext von GIS zu den Netzwerkanalysen. Im Fall dieser Arbeit handelt es sich um ein reines Netzwerk. Dabei liegt im Gegenteil zu einem Flussnetzwerk kein gerichteter Fluss innerhalb des Graphen vor, sondern es besteht nur aus einer Topologie und einer Konnektivität. Im Rahmen eines GIS wird ein Netzwerk-Daten-Modell erstellt, das aus zwei Hauptkomponenten besteht: den Knoten und die Kanten. Ein Knoten entsteht immer dann, wenn sich zwei oder mehr Kanten kreuzen. (Curtin 2007) Ein Einzugsgebiet wird rund um einen POI errechnet. (Abbildung 11) Es umfasst alle umgebenden und erreichbaren Punkte in einem Netzwerk (z.B.: Straßenabschnitte) innerhalb einer bestimmten Impedanz. Ist die Impedanz als Zeit angegeben so zählt die schnellste Route. Ist der Weg als Impedanz festgelegt, so gilt die kürzeste Route. Beide Arten haben gemein, dass sie die kostengünstigste Route als Ziel haben. (Cullinan, Hynes & O'Donoghue 2008) Für diese Arbeit werden als Impedanzen sowohl die Zeit als auch der Weg herangezogen – beide zusammen ergeben die geforderten Kosten. Die Berechnung der Reisekosten kann mit folgender Formel 2 dargestellt werden: (Talen & Anselin 1998)

$$Z_i^T = \sum_j d_{ij} \tag{2}$$

Z_i^T…Reisekosten für einen Punkt i

d_{ij}…Distanz zwischen einem Ursprungspunkt i und einer Einrichtung j

i…vom Nutzer genannter POI

j…die Zieleinrichtung ergibt sich durch die angegebene Impedanz – deshalb ist es ein fiktiver Punkt innerhalb des Netzwerkes

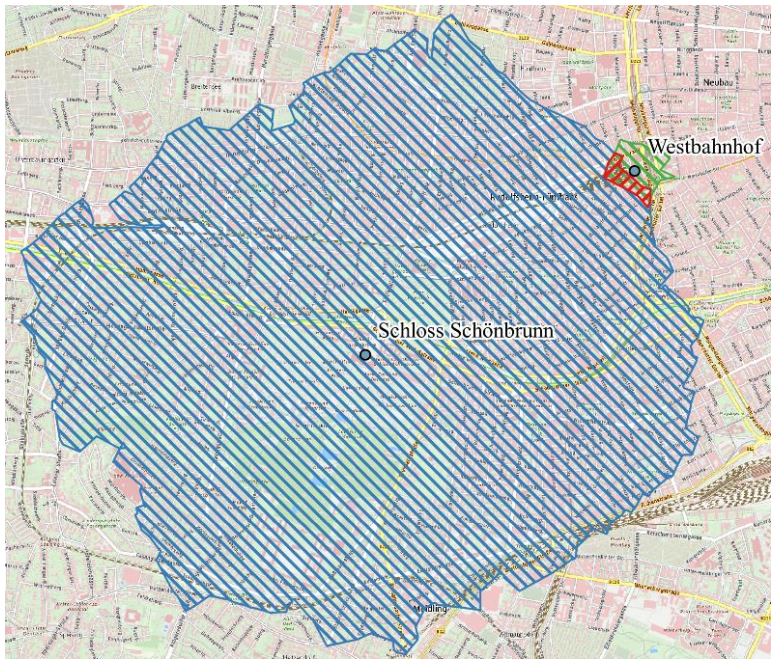


Abbildung 11 Beispielhafte Einzugsgebietsanalyse zweier Standorte (grüne und blaue Fläche) und deren Überschneidungsbereich (rote Fläche) (Hintergrund: basemap.at)

Algorithmen zur Suche nach dem kürzesten Weg sind bereits in die GI-Systeme, wie ESRI ArcMap oder QGIS implementiert. Als Minimalanforderung ist ein routingfähiges Netzwerk, mit hinterlegten Längen und Maximalgeschwindigkeiten essenziell. Mögliche Einschränkungen, wie Fahr- oder Abbiegeverbote sind optional. (Shahid et al. 2009) Der wohl bekannteste Algorithmus zum Finden des kürzesten Weges ist der Dijkstra-Algorithmus. Dieser arbeitet sich von Knoten zu Knoten zum Ziel und geht dabei immer den Weg des geringsten Gewichts (Impedanz). Dieser Algorithmus funktioniert nur bei rein positiven Kantengewichten. (Blum 2004: 100)

2.7 Zwischenfazit

Das gesamte Kapitel 2 gibt einen Überblick über die berücksichtigte Terminologie in dieser Arbeit. Es zeigt sich, dass die theoretischen Grundlagen zu der sogenannten räumlichen, natürlichen Sprachprozessierung vorhanden sind. Die in den meisten Studien

analysierte Datenbasis stammt entweder von Twitter oder aus Zeitungsartikeln. Der Fokus bei der inhaltlichen Auswertung liegt entweder auf einer widerspruchsfreien Geokodierung der erkannten Entitäten und/oder auf der Interpretation von Präposition innerhalb eines Textes mit geographischem Bezug.

Für die praktische Auswertung ergibt sich der in Abbildung 12 visualisierte Prozess, der für alle Geo-text Daten angewandt werden kann. In einem ersten Schritt gilt es die geographisch relevante Information aus einem Konvolut an Daten zu erkennen und zu extrahieren. Weiters soll mittels Geoparsing, den erkannten Toponymen ein räumlicher Bezug zugewiesen werden. Der dritte Schritt wird von Hu 2018 als Schlüsselstelle beschrieben, da dieser mannigfaltige Methoden zur Analyse beinhalten kann. Dies reicht von der Sentiment-Analyse über die räumliche Autokorrelation bis hin zur computerunterstützen Interpretation von Präpositionen. Die Evaluierung soll einen kritischen vorletzten Schritt darstellen, damit feststellbar ist, ob das neue Wissen auch valide ist. Eine qualitative Evaluierung soll vor allem dazu dienen, intuitives Verständnis zu prüfen. Bei der Quantität zählen nur die Metriken, welche die Fehlerraten oder die Genauigkeiten widerspiegeln können. Den Abschluss bilden Visualisierungen – in dieser Arbeit besonders die Netzwerk-Visualisierungen. (Hu 2018)

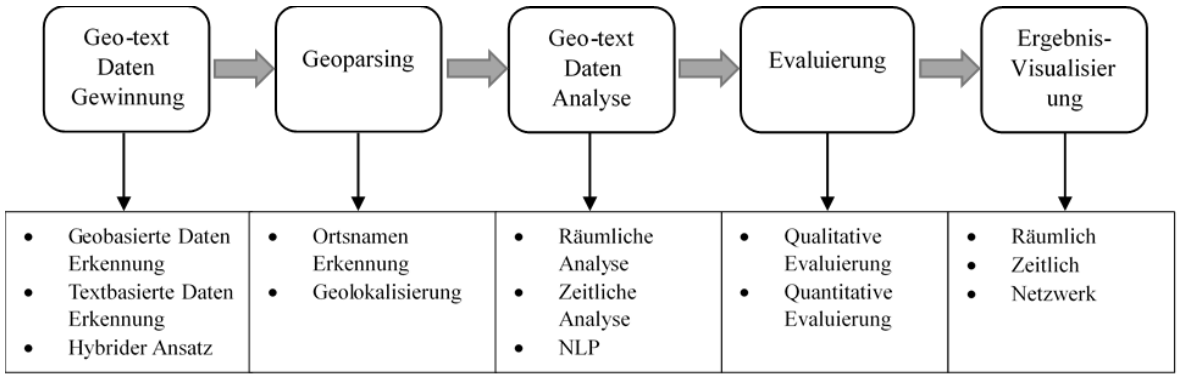


Abbildung 12 Prozess beim Analysieren von Geo-text Daten (nach Hu 2018: 13)

3 Datenquellen

Dieser Abschnitt soll einen Überblick über die verwendeten Datenquellen liefern. Die einzelnen Anbieter haben meist mehr Attribute, als die Verwendeten veröffentlicht, jedoch sind diese hier nicht von Relevanz. Zu Beginn werden allgemeine Fakten der Vermittlungsplattform Airbnb erläutert. Der Datenanbieter für Airbnb ist Inside Airbnb. Hier ist ein Großteil der Daten unverändert und allesamt maschinenlesbarer aufbereitet. Im Anschluss daran wird in weiteren Unterkapitel die österreichweit verfügbare Graphenintegrations-Plattform (GIP) und die POIs der Stadt Wien, erläutert. Auf der Internetseite *data.gv.at* werden die dafür notwendigen Pfade zur Verfügung gestellt. Sie stehen unter der *creative commons* Lizenz (Typ: CC BY 3.0 AT) Das bedeutet, dass die Daten geteilt und unbegrenzt, auch für kommerzielle Zwecke, bearbeitet werden dürfen, sofern immer eine Namensnennung gegeben ist.

3.1 Airbnb

Airbnb ist eine US-amerikanische Firma, die auf Webseite *airbnb.com*, Privatunterkünfte von und für Privatpersonen vermittelt. Das Unternehmen wurde 2008 im Silicon Valley in Kalifornien, USA gegründet. Der Wert des Unternehmens liegt zurzeit bei zirka 38 Milliarden US-Dollar. (Forbes 2018) Seit dem Gründungstag wurden mehr als 400 Millionen Gästeankünfte in mehr als 191 Länder gezählt. (Airbnb 2019b) Das Wort „Airbnb“ steht nicht nur als Name für ein Unternehmen, sondern wird in dieser Arbeit auch als Synonym für alle Ausprägungen von „Unterkunft“ verwendet.

Die Idee hinter Airbnb ist, dass jene Personen, welche über eine private Unterkunft verfügen, die sie für sich selbst nicht benötigen, vermieten. Als Vorteil dieses Systems sieht Airbnb, dass es Touristen ermöglicht wird, so zu leben wie die ansässige Bevölkerung, sei es durch die regionale Einrichtung oder aber auch den Standort in nicht-touristischen Stadtvierteln. Außerdem ist es meist günstiger als die Übernachtung in einem Hotel, da Gebühren, welche Hotels abführen müssen, nicht oder nur teilweise geltend gemacht werden. Airbnb verdient dabei an jeder Transaktion, die zwischen Gast und Gastgeber durchgeführt wird. (Airbnb 2019a) Die Hosts zahlen drei Prozent der pro-Nacht-Rate einer Buchung an Airbnb. Die Gäste müssen eine Servicegebühr zwischen sechs und zwölf Prozent an Airbnb entrichten. Außerdem sind die Gastgeber für die Einhaltung der örtlichen Gesetze verantwortlich. (Henten & Windekilde 2016)

Für Wien sind zurzeit 10.714 Unterkünfte online. Davon sind 72,4% als „gesamte Unterkunft“, 26,6% als „Privatzimmer“ und 1% als „geteilte Zimmer“, deklariert. Laut der Plattform *Inside Airbnb* weisen 57,8% der Unterkünfte eine hohe Verfügbarkeit auf und sind theoretisch das ganze Jahr über Verfügbar. Außerdem haben 44,5% der Gastgeber mehr als ein Quartier im Angebot. (Cox 2018a) In Abbildung 13 ist die räumliche Verteilung der Unterkünfte in quantitativer Form, sowie als Verhältnis von Airbnbs zu Wohnungen dargestellt. Es zeigt sich, dass vor allem die inneren Bezirke von Wien eine

hohe Anzahl an Airbnbs aufweisen. Im Stadtzentrum, dem 1. Bezirk werden 7,28% aller bewohnten Wohnungen zurzeit als Airbnbs genutzt bzw. wurden sie zumindest einmal als solche genutzt. Es ist nicht unterschieden worden, ob die Unterkünfte noch buchbar sind – im Datenbestand sind sie dennoch vorhanden. Für diese Berechnungen wurde die Anzahl an Airbnbs durch die Anzahl an bewohnten Wohnungen je Bezirk dividiert. Die Zahlen zu den Wohnungen in Wien stammen von der MA23, der Magistratsabteilung für Wirtschaft, Arbeit und Statistik.

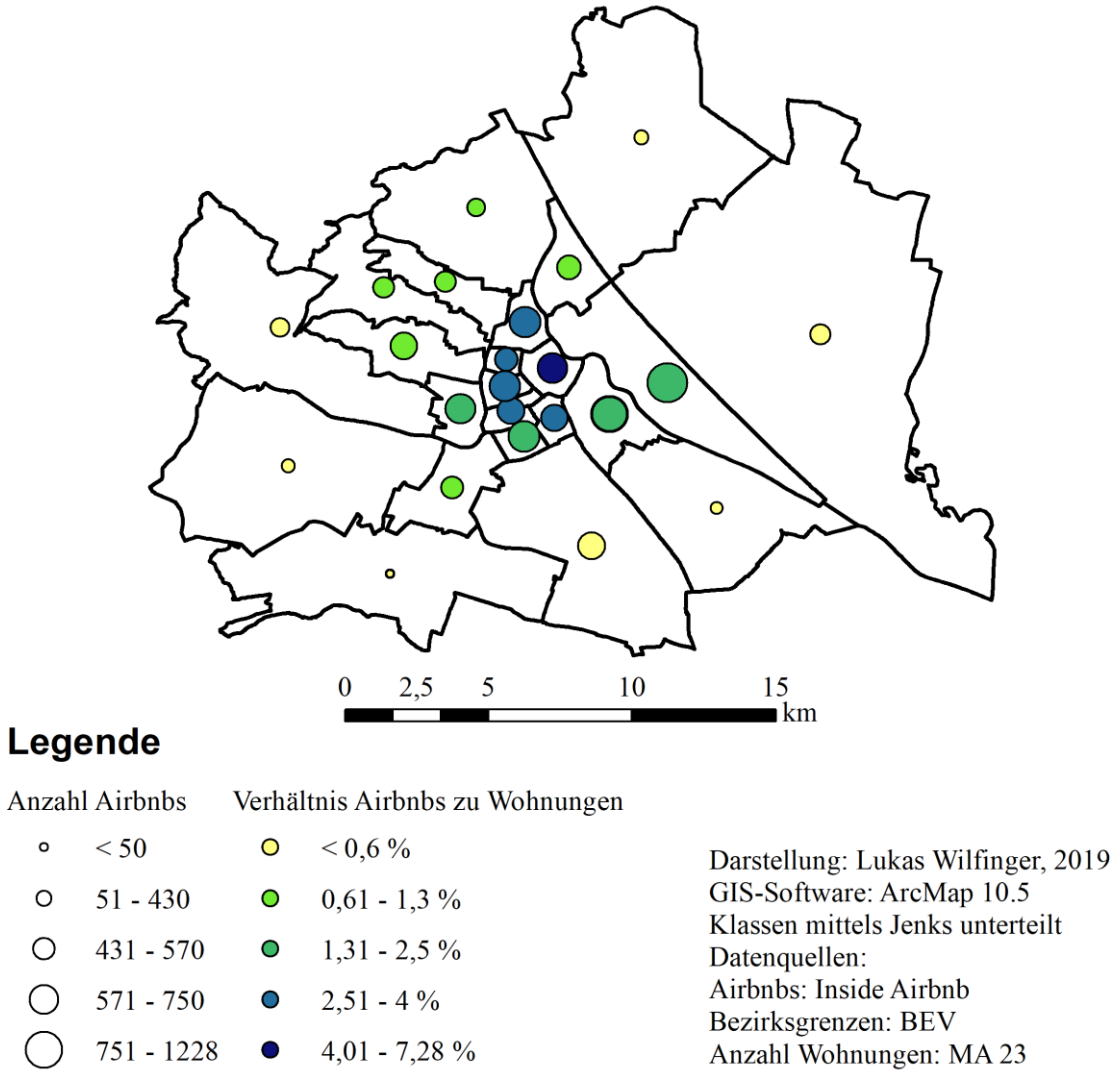


Abbildung 13 Anzahl Airbnbs und Verhältnis von Airbnbs zu Wohnungen je Bezirk

3.2 Inside Airbnb

Die Onlineplattform Inside Airbnb besteht, laut eigener Beschreibung auf der Homepage, *aus einer unabhängigen und nicht-kommerziellen Reihe an Werkzeugen, die dabei helfen sollen herauszufinden, wie Airbnb in Städten rund um die Erde verwendet wird.* Hinter der Webseite stehen der Kommunikationsaktivist und Technologe Murray Cox und der

Grafikdesigner John Morris. Inside Airbnb stellt die Daten auf neutrale Art und Weise dar, ohne wertende Begriffe zu verwenden. Außerdem entspricht ein Großteil an herunterladbaren Daten, jenen Informationen, die auch auf der Webseite Airbnb sichtbar sind – nur eben in textueller Form. Nichtsdestotrotz versucht Inside Airbnb durch die Bereitstellung der Daten investigative Analysen zu unterstützen. So wird beispielsweise kritisiert, *dass Airbnbs immer mehr als permanent verfügbare Unterkünfte vermietet werden*. Dieser Umstand sorgt in vielen europäischen Großstädten für erhöhte Mietpreise am Wohnungsmarkt und sozialen Unruhen. (Cox 2018a)

3.2.1 Rechtliche Grundlage

Über die Nutzungsbedingungen der Webseite von Airbnb ist in Erfahrung zu bringen, dass es verboten ist *Robots, Spider, Crawler, Scraper oder sonstige automatisierte Mittel oder Verfahren zu nutzen um auf die Airbnb-Plattform zuzugreifen, von der Airbnb-Plattform Daten oder sonstige Inhalte zu ziehen oder auf andere Weise mit ihr zu interagieren*. (Airbnb 2019c) Mit welcher Methodik die Daten durch Inside Airbnb zusammengetragen wurden, ist nicht bekannt. Auf der Internetseite wird nur auf die verwendeten Technologien eingegangen. (Cox 2018b) Ist allerdings eine der oben genannten Verfahren zum Einsatz gekommen, dann scheint es, als wird dieses Vorgehen von Airbnb toleriert, sofern gewisse ‚Crawler-Etiquetten' (Airbnb 2019d) eingehalten werden, wie Ke 2017 zeigte. Die empfindlichsten Daten von Airbnb sind Informationen zur Auslastung einzelner Unterkünfte und deren Typs (gesamte Unterkunft, eigenes Zimmer oder geteiltes Zimmer). Diese sind aufgrund eines möglichen Steueraufkommens vor allem für lokale Regierungen von großem Interesse. (Martineau 2019) Die Analysen, die sich dahingehend auf die Daten von Inside Airbnb stützen wurden von Airbnb als ‚ungenau' (engl. *inaccurate*) tituliert. (Hutchins & Mahoney 2015) Die für diese Arbeit verwendeten Daten sind nur die jeweiligen Beschreibungen, Bewertungen und Koordinaten zu den einzelnen Unterkünften. Es werden nur öffentliche und keine personenbezogenen Informationen des Verfassers oder der Verfasserin analysiert. Außerdem werden nur Daten verwendet, welche nicht, wie die Preise pro Nacht oder die Belegung der Unterkunft über das Jahr hochgerechnet wurden. Wenn diese Art an Daten untersucht wird, die auch öffentlich zugänglich ist, so verweisen Cheng & Jin 2019 daraufhin, dass überhaupt kein ethisches Augenmerk auf die Analyse zu legen ist. Nichtsdestotrotz wird in Abschnitt 4.4.3 auf das ethische Verantwortungsbewusstsein eingegangen, damit alle Aspekte einer Datenanalyse abgedeckt sind.

3.2.2 Aufbereitung der Daten

Die zur Verfügung stehenden Daten stellen für jede Stadt nur einen zeitlichen Schnappschuss dar. Für diese Arbeit werden die Daten über Wien mit dem Zeitstempel vom 15.11.2018 verwendet. Die Informationen sind als Text-Dateien im *txt*-Format herunterzuladen. In Tabelle 7 sind die relevantesten Dateien und die darin enthaltenen Attribute und deren Typ aufgelistet.

Name	Variablen	Typ	Beschreibung
Listings	ID	Nummer	ID des Airbnbs
	Name	Nummer & Text	Name der Unterkunft
	Host_ID	Nummer	Nummer des Hosts
	Host_name	Text	Name des Hosts
	neighbourhood	Text	Bezirksname
	Latitude	Nummer	Breitengrad
	Longitude	Nummer	Längengrad
	Room_type	Text	Art der Unterkunft
	Price	Nummer	geschätzter Preis
	minimum_nights	Nummer	Mindestanzahl an zu buchenden Nächten
	number of reviews	Nummer	Anzahl an Bewertungen
	last_review	Datum	Tag der letzten Bewertung
	reviews_per_month	Nummer	Anzahl an Bewertungen pro Monat
	calculated_host_listings_count	Nummer	Anzahl an Airbnb des Hosts
	availability_365	Nummer	Verfügbare Tage im Jahr
Reviews	listing_id	Nummer	ID des Airbnbs
	ID	Nummer	ID der Bewertung
	date	Datum	Datum der Bewertung
	reviewer_name	Text	Name des Bewerters/Gastes
	comments	Text	Inhalt des Kommentars

Quelle: Inside Airbnb, 2019

Tabelle 7 Beschreibung der verwendeten Daten von Inside Airbnb

Als Verbindungselement gilt die *listing_id*. Mit Hilfe dieser kann festgestellt werden welche Bewertung sich auf welche Unterkunft bezieht. Bei einer Orientierung an Tabelle 7 kann für die vorliegende Information die folgende Einteilung in die jeweiligen Datentypen erfolgen. Die textuelle Information in den einzelnen Spalten ist unstrukturiert, da es sich um reinen Text handelt. Der Aufbau der gesamten Datei ist allerdings als Tabelle gedacht und kann ohne weiteres in eine Datenbank implementiert und über den Schlüssel mit anderen Informationen verknüpft werden. Dies spricht für einen strukturierten Datentyp.

3.2.3 Qualität von nutzergenerierten Informationen

Dieses Kapitel wurde bis einschließlich Abbildung 14 eins-zu-eins aus dem Abschnitt *Qualität, Fehler, Merkmale und Anforderungen* der Masterarbeit *Bildung einer Heuristik zur Plausibilisierung und Validierung von Verkehrsmanagementdaten* übernommen. (Wilfinger 2018) Das hat vor allem den Hintergrund, da sich die Ansichten zu Qualität im Allgemeinen und deren Bestandteile nicht verändert haben. So ist zum Beispiel die angegebene DIN EN ISO 9000:2015, noch immer die aktuellste Auflage.

Das Wort „Qualität“ leitet sich vom lateinischen Wort *qualitas* ab und bedeutet so viele wie Beschaffenheit oder Eigenschaft. (Brockhaus 2018) Das Interessante daran ist, dass Qualität meist positiv assoziiert wird, obwohl die eigentliche Bedeutung keine Wertung enthält. Dies rührt daher, dass es oftmals mit hohen Leistungsmerkmalen in Verbindung gebracht wird. Aber eigentlich ist der Begriff nach obiger Bedeutung als neutral anzusehen. (Jentsch 2009: 6)

Die Veränderung der Definition von Qualität lässt sich beim Vergleich zweier Versionen nach DIN-Norm erläutern. Heißt es in der DIN EN ISO 8402:1995: „*Qualität ist die Gesamtheit der Merkmale, die eine Einheit zur Erfüllung vorgegebener Forderungen geeignet macht.*“ So änderte sich die Definition in der aktuellen DIN EN ISO 9000:2015-11 auf „*Grad, in dem ein Satz inhärenter Merkmale eines Objekts Anforderungen erfüllt*.“ Es zeigt sich eine Verschiebung des Hauptaugenmerks von den Merkmalen hin zu den Anforderungen eines Objektes. Qualität ist nicht messbar und auch nicht bivalent. (Brüggemann & Bremer 2015: 4) Nichtsdestotrotz ist die Bestimmung der Qualität ein Soll-Ist-Vergleich. Jede Abweichung vom Soll wird als Fehler vermerkt und vermindert somit die Güte. (Reinhart, Lindemann & Heinzl 1996: 6) Ein Fehler wiederum ist dann lt. DIN EN ISO 9000:2015 die „*Nichterfüllung einer Anforderung*“. Im Gegensatz dazu sind Mängel, in etwas abgeschwächter Form zu Fehler, „die Nichterfüllung einer Anforderung in Bezug auf einen beabsichtigten oder festgelegten Gebrauch.“ (Jentsch 2009: 7)

In der aktuellen Definition nach ISO 9000 gibt es zwei weitere Begriffe die es, mit dem Blick auf den Forschungsgegenstand, wert sind noch näher zu betrachten. Das sind zum einen die Merkmale und zum anderen die Anforderungen.

Merkmale

Eine kennzeichnende Eigenschaft wird Merkmal genannt. Soll diese, einer bestimmten Anforderung unterliegen, so handelt es sich um ein Qualitätsmerkmal. Die Untersuchung, ob diese eingehalten werden, ist ein zentraler Bestandteil vom Qualitätsmanagement. Nicht alle Qualitätsmerkmale sind für den Kunden von gleicher Wichtigkeit bzw. lösen eine gleiche Emotion aus – wie das Kano-Modell zeigt (Abbildung 14). (Jentsch 2009: 7)

Anforderungen

Wenn Anforderungen an das Objekt gestellt werden, ist es möglich dessen Erfüllungen oder Nichterfüllung festzustellen. Nur dadurch ist ein Urteil über die Qualität gegeben. (Reinhart, Lindemann & Heinzl 1996: 6) Anforderungen sind nach DIN EN ISO 9000:2015 als „festgelegte, üblicherweise vorausgesetzte und verpflichtende *Erfordernisse oder Erwartungen*“ definiert. Das oben bereits erwähnte Kano-Modell unterscheidet zwischen drei unterschiedlichen Anforderungen: Basis-(„must-be“), Leistungs- („one-dimensional“) und Begeisterungsanforderungen („attractive“). Damit wird der „Zusammenhang zwischen der Erwartungserfüllung und den Verläufen der Zufriedenheitsfaktoren“ grafisch dargestellt. (Hölzing 2008: 85)

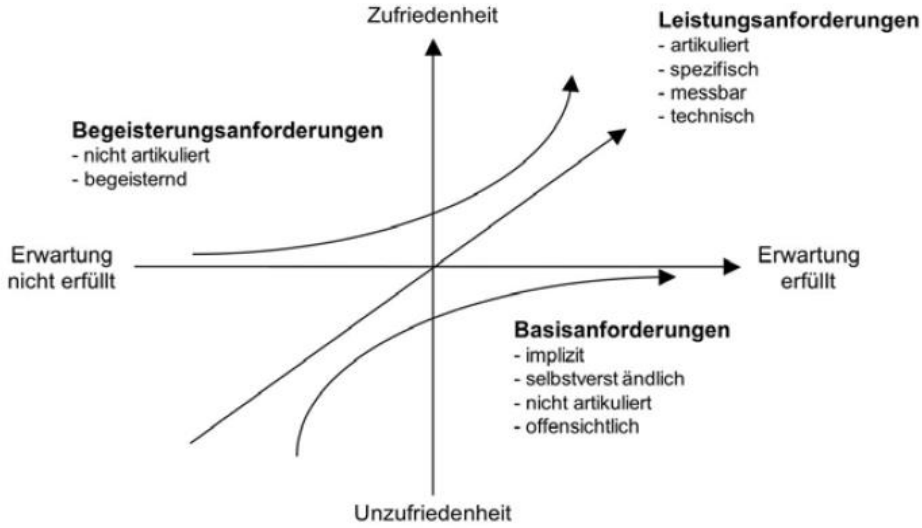


Abbildung 14 Kano-Modell (Hölzing 2008: 85)

Damit die Datenqualität im Allgemeinen bestimmt werden kann, bedarf es einer Festlegung der Anforderungen. Dabei sind eine wesentliche Hypothese und eine themenbezogene Problemstellung zu berücksichtigen. Im folgenden Absatz wird eine Hypothese zu nutzergeneriertem Inhalt definiert, die durch den zweiten Absatz bekräftigt werden soll.

Es gibt allgemeines Interesse an nutzergenierten Daten. (Krumm, Davies & Narayanaswami 2008) Es handelt sich bei solchen Daten um Informationen, die von Personen aus der „gewöhnlichen“ Öffentlichkeit stammen und an dieser auch aktiv teilnehmen. (Foren, soziale Medien, Blogs, etc.) Damit geht einerseits die Kontrolle über den Inhalt verloren und obendrein ist es nicht möglich festzustellen inwieweit der erstellte Text eine Relevanz für das bestimmte Thema aufweist. In diesem Zusammenhang wird von einer potenziell niedrigen Informationsqualität aus einer Masse diskutiert. (Lukyanenko 2014)

Die Herausforderung aus der obigen Hypothese, dass nutzergenerierter Inhalt nicht kontrolliert werden kann, wird noch durch die allgemeinen Geschäftsbedingungen von Airbnb erhöht. Darin ist enthalten, dass die Bewertungen von Gästen oder Gastgebern nicht auf deren Richtigkeit überprüft werden. Es ist festgeschrieben, dass der Inhalt nicht auf dessen Rechtmäßigkeit geprüft wird. Eine Rechtschreibüberprüfung oder eine Kontrolle der Morphologie des Textes kann daraufhin nicht garantiert werden. Kontrolliert werden nur personenbezogene Bewertungen von Gästen oder Gastgebern. Diese dürfen nicht beleidigend oder diffamierend sein. (Airbnb 2019e)

3.2.4 Festlegung der Anforderungen

Ein effizientes Arbeiten mit nutzergenerierten Daten erfordert ein vorheriges Festlegen von Anforderungen. Diese steuern, welche Unterkünfte und Sätze in die Analyse

miteinfließen dürfen. Mit dieser Vorgehensweise wird verhindert, dass falsch kodierte Wörter, räumlich nicht relevante Sätze und morphologisch inkorrekte Sätze in den Auswertungsprozess kommen. Fürs erste gilt es die Basis- und Leistungsanforderungen an die Airbnb-Beschreibungen und -Bewertungen festzulegen.

- Basisanforderungen:
  - Der von Inside Airbnb herunterzuladende Datensatz muss maschinenlesbar sein.
  - Die Datensätze müssen auf Unterkünfte und deren Beschreibungen und Bewertungen verweisen.

Das sind all jene Anforderungen, die der Analyst implizit an die Daten stellt, ohne sich zu sehr auf die vorgegebene Thematik zu beschränken. Wenn die Basisanforderungen erfüllt sind, stellt sich beim Empfänger noch keine Zufriedenheit ein, jedoch ist bei Nichterfüllung eine weitere Analyse nicht möglich. Im Gegensatz dazu haben die Leistungsanforderung einen negativen, als auch einen positiven Einfluss auf die Zufriedenheit. (Hölzing 2008)

- Leistungsanforderungen:
  - Die Angabe des Standorts der Unterkunft muss als X- und Y-Koordinate gegeben sein.
  - Das Airbnb verfügt über eine Beschreibung oder es gibt zumindest eine Bewertung, um die räumlichen Beziehungen zu validieren.
  - Die Beschreibung und die Bewertung sind in englischer oder deutscher Sprache gehalten. Andere Sprachen oder Kodierungsfehler können nicht berücksichtigt werden.
  - In der Bewertung oder der Beschreibung ist ein Hinweis über die räumlichen Nachbarschaftsbeziehungen gegeben. Diese sind mittels zeitlichen Distanzangaben oder Angaben in Meter beschrieben.
  - Bei der Position der Unterkunft muss es sich um den realen Standort handeln. Eine Auswertung von verschleierten Unterkünften ist nicht das Ziel dieser Arbeit.

Zum Schluss kann noch die Begeisterungsanforderung erfüllt sein. Diese ist nicht zwingend erforderlich – erleichtert jedoch die Auswertung. Solche Anforderungen können nur einen positiven Einfluss auf die Zufriedenheit haben. (Hölzing 2008)

- Begeisterungsanforderung:
  - In mindestens einer Bewertung befindet sich einmal ein Synonym für „Airbnb“, eine Distanzangabe und ein POI. Zu den Synonymen zählen unter anderem: Wohnung, Unterkunft oder Haus.

3.3 Graphenintegrations-Plattform

Im Zuge der Graphenintegrations-Plattform (GIP) wurde ein aktueller österreichweit verfügbarer Straßengraph erstellt. Die Partner dieses Projektes waren die neun österreichischen Bundesländer, sowie die ASFINAG, die ÖBB und das BMVIT. In einem mehrjährigen Projekt wurden an alle Gemeinde in Österreich topographische Karten mit den derzeitigen Straßengraphen gesendet. Der Prozess sah vor, dass Fehler vom dortigen Sachbearbeiter ausgebessert, danach wieder zurückgesendet und in ein GIS eingearbeitet werden. Um die finanziellen Dimensionen des damaligen Projekts aufzuzeigen folgt eine Aufstellung für das Bundesland Niederösterreich. In das damalige Projekt des niederösterreichischen Verkehrsdatenverbunds (GIP.nö) wurden 1,8 Millionen Euro über eine Dauer von zwei Jahren investiert. Dabei sind alle Gemeindestraßen (fahrstreifengenau), Güterwege, Brücken, Tunnel, Kreuzungen mit Abbiegerelationen, Rad- und Fußgängerinfrastruktur, Bushaltestellen, Nebenbahnen (NÖVOG-Strecken) und Zugangswege zu Bahnhöfen erfasst worden. Das waren in Summe 80.000 km Straßen, 70.000 Hausnummern und 13.000 Haltestellen. (GIP 2018) Seit Ende dieses Projektes steht ein österreichweiter und frei verfügbarer Straßengraph zur Verfügung, der auch von Ministerien eingesetzt wird.

Das Netzwerk für Wien enthält alle Autostraßen, U- und Straßenbahnlinien, sowie Geh- und Radwege. Pro Kante sind auch immer etwaige Restriktionen festgehalten. Beispiele hierfür sind Einbahnen, Maximalgeschwindigkeiten und Abbiegeverbote. Damit ein Routing zwischen der GIP und den POIs möglich ist, muss eine dazugehörige Verbindungskante händisch digitalisiert werden.

Link zur Ressource:

http://www.gip.gv.at/ogd-228.html (abgerufen am 12.02.2019)

3.4 Open Data der Stadt Wien

Die POIs für diese Arbeit bestehen zum einen aus den touristischen Hotspots in Wien, als auch aus U-, Schnell- und Straßenbahnhaltestellen. Beide Datensätze werden von der Stadt Wien inklusive Verortung zur Verfügung gestellt. Für eine spätere Verfeinerung des Ergebnisses wird noch auf die flächenhafte Realnutzung der Stadt Wien zurückgegriffen. All diese Daten sind frei verfügbar und dürfen auch geteilt und verarbeitet werden. Das globale Referenzsystem WGS 84 liegt diesen räumlichen Daten zugrunde.

Link zu den Haltestellen: (abgerufen am 12.02.2019)

https://www.data.gv.at/katalog/dataset/stadt-wien_haltestellenstandortewien

Link zu den Sehenswürdigkeiten: (abgerufen am 12.02.2019)

https://www.data.gv.at/katalog/dataset/f4e80988-c139-4953-8176-b3d6d03f6449

Link zur Realnutzung: (abgerufen am 02.06.2019)

https://www.data.gv.at/katalog/dataset/stadt-wien_realnutzungskartierungab200708wien

Die Informationen sind allesamt über die Seite opendata.gv.at abrufbar. Die Plattform veröffentlich mehrheitlich offene Verwaltungsdaten (OGD). *Dies sind nicht personenbezogene Informationen, die von öffentlichen Stellen gesammelt, erstellt oder bezahlt wurden und freiwillig sowie kostenlos der Allgemeinheit zur Verfügung gestellt werden.* (Data.gv.at 2019)

3.5 Frei verfügbare Informationen und Technologien

Sowohl die GIP als auch die POIs der Stadt Wien werden kostenlos im Internet zur Verfügung gestellt. Ein Grund dafür ist, dass Österreich ein Partner der Open Government Partnership (OGP) ist. Diese garantiert eine Verbindlichkeit, welcher die nationalen Regierungen nachkommen sollen. Es wird Transparenz gefördert, der Bürger gestärkt und außerdem werden Technologien zur Stärkung der Regierungsführung gefunden. Es gibt vier Bereiche in denen die Partnerländer deren Motivation zur Teilnahme an der OGP zeigen sollen: (Attard 2016)

- Steuerliche Transparenz
- Leichter Zugang zu Information
- Einkommens- und Vermögensangaben
- Engagement der Bürger

Die Gründe, weshalb Daten von Gemeinden, Ländern oder Staaten überhaupt veröffentlich werden, sind das Zeigen von (1) Transparenz und (2) Verantwortlichkeit, (3) sowie die Bereitschaft die Demokratie zu stärken. Der erste Punkt ermöglicht zum einen Feedback, aber auch eine Überwachung von Regierungsaktivitäten. (Yang et al. 2013) Die Verantwortlichkeit besteht in diesem Zusammenhang aus zwei Komponenten. Der Bereitstellung der Information selbst und der Rechtfertigung dieser. Durch die Veröffentlichung für die Bürger wird die Rechtfertigung auf die Probe gestellt. (Lourenço 2013) Die Demokratie wird vor allem dann gestärkt, wenn Zahlen über das Budget in regelmäßigen Abständen veröffentlicht werden. Der Bürger bekommt Einblick in den Staatshaushalt und erkennt Missstände – somit ist bevölkerungszentrierte Politik möglich. (Mutuku & Colaco 2012)

Bei den hier verwendeten Datenquellen ist ein Sprechen von „Stärkung der Demokratie“ zu weit gegriffen. Die Transparenz und eine Stärkung der Fähigkeiten der Bürger stehen im Vordergrund. Mit der Veröffentlichung ist es Schülerinnen und Schülern, Studierenden oder Privatpersonen, ohne der Tätigung von Zahlungen möglich, auf Daten zurückzugreifen, die auch von behördlicher Seite verwendet werden.

4 Methodik zur Standortermittlung

Dieses Kapitel soll die wichtigsten theoretischen und praktischen Überlegungen zur Umsetzung der Analyse beinhalten. Begonnen bei der Beschreibung des gesamten praktischen Aufarbeitungsprozesses, über die Erläuterung welche Technologien für die Auswertung verwendet wurden, bis hin zum ethischen Grundgedanken bei solchen Analysen.

4.1 Vorgehensweise

Die Aufarbeitung der praktischen Fragestellung soll logisch nachvollziehbar sein. Dieser Abschnitt soll klären wie Umsetzung erfolgt und welche Überlegungen dahintergestanden sind.

Als Datenquelle wurde bereits Inside Airbnb bestimmt. Das Untersuchungsgebiet wurde mit Wien auch schon eingeschränkt. Nähere Informationen sind dazu aus Abschnitt 1 zu entnehmen. Es gilt nun aus diesem Konvolut an Unterkünfte eine geeignete und repräsentative Stichprobe zu entnehmen.

Was nach all diesen Vorselektionen aus den Ausgangsdaten bleibt, sind textuelle Daten, die in einem relationalen Schema vorliegen. Innerhalb dieses Schemas gibt es Spalten, die für die Analyse nicht von Notwendigkeit sind – das gleiche gilt auch für die Daten von *opendata*. Solche Spalten sind zu entfernen, um unnötige Interaktionen zu vermeiden und eine Aufgabenangemessenheit zu garantieren. Das Entfernen von Spalten ist nur ein Beispiel für die Aufgaben, die in Kapitel 4.3 zu erledigen sind. Es folgt der aufwendigste Teil dieser Arbeit, die Extraktion von räumlich relevanter Information aus den textuellen Daten. Begonnen wird mit der Tokenisierung und damit der Teilung der Sätze in einzelne Wörter. Es folgt eine Suche nach Wort- und Zahlenfolgen, die auf eine räumliche Relation hindeuten. Das bedeutet, dass Annahmen zu jenen Wortfolgen getroffen werden müssen, welche einen Rückschluss auf räumliche Beziehungen zulassen. Diese Information ist in einem zusätzlichen Schritt zu extrahieren, auch wenn teilweise irrelevante Daten mitenthalten sind, beispielsweise wenn die Annahmen nicht vollständig waren bzw. es mit dieser Vorgehensweise nicht abbildbar war. Es folgt die namentliche Objekterkennung. Diese bedarf zunächst einer Normalisierung der Daten, die in diesem Fall händisch durchgeführt wurde. Die Gründe für diese Vorgehensweise sind in Kapitel 4.4.2 angeführt. Ein harter Abgleich zwischen den normalisierten POIs aus den Airbnb-Daten und den OGD-Daten der Stadt Wien, dient als Anreicherung mit Koordinaten. Die jeweilige räumliche Relevanz wird dann wiederum händisch festgelegt. Durch eine solche Kodierung kann eine weitere effiziente Verarbeitung im GIS durchgeführt werden. Mit dieser Technologie ist es gegen Ende der praktischen Ausarbeitung möglich, die Daten zu visualisieren, räumlich zu analysieren und Verbesserungsmöglichkeiten zu finden. Der soeben beschriebene Ablauf ist in Abbildung 15 dargestellt.

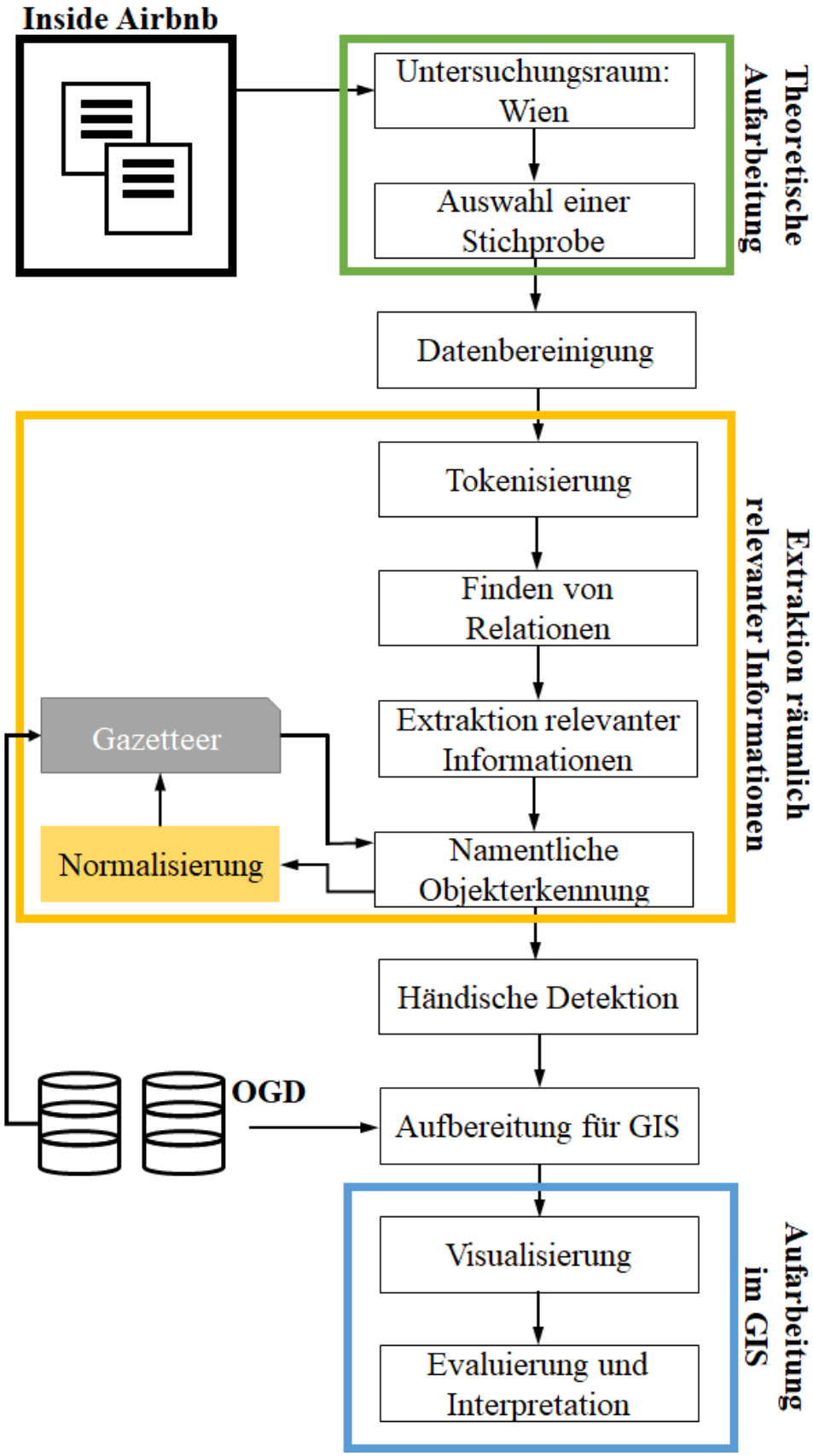


Abbildung 15 Ablauf der praktischen Analyse

4.2 Untersuchungsraum

Als Untersuchungsraum, auf den sich die Analysen beschränken, wurde Wien ausgewählt. Für ein Testen der Methodik werden nicht alle Airbnbs im Stadtgebiet analysiert, sondern nur jene, die auch den Anforderungen aus Abschnitt 3.2.4 entsprechen. Wichtig vor allem ist, dass der reale Standort bekannt ist. Die Daten von Inside Airbnb liefern zwar eine Variable mit der angegeben wird, ob der Standort exakt oder ungenau ist, allerdings hat eine händische Probe von 20 zufällig ausgewählten Unterkünften deren reale Position nicht bestätigen können. Als wahrer Standort wurde die Angabe auf der Internetseite festgelegt. Bei allen, angegebenen exakten Standorten, handelte es sich um „verschleierte" Positionen. Die Methode, wie Inside Airbnb auf den genauen Standort zu schließen glaubt, ist nicht bekannt. Für die Analyse müssen jedoch exakte Standortangaben gefunden werden. Somit ist von nun an, die Angabe auf der Internetseite von Airbnb zur jeweiligen Unterkunft als korrekte Position anzusehen.

Damit eine Größenordnung von verschleierten zu exakten Positionsangaben auf *airbnb.com* vorstellbar ist, wurden 200 Airbnbs in Wien mittels Zufallsstichprobe untersucht und dabei dreizehn Kartendarstellungen mit der Überschrift „*es handelt sich um den genauen Standort der Unterkunft"* gefunden. In Wien gibt es zurzeit 11.276 aktive Airbnbs (Stand: 05.03.2019). Alle Unterkünfte händisch auf deren realen Standort zu untersuchen würde zu viel Zeit in Anspruch annehmen. Es wird nach den obigen Zahlen angenommen, dass bei zirka fünf Prozent aller Airbnbs in Wien der Standort exakt angegeben ist. Für die Methodenentwicklung steht demnach eine Grundgesamtheit von etwa 564 Unterkünften (Basis: 11.276 Unterkünfte und davon fünf Prozent) zur Verfügung. Daraus soll nun, wissenschaftlich nachvollziehbar, eine Stichprobe gezogen werden, um die Methode zu eruieren. Das folgende Unterkapitel soll klären, wie hoch die Anzahl an zu untersuchenden Unterkünften mindestens sein muss, um die darauffolgenden Methodenentwicklung zu bestätigen.

4.2.1 Auswahl einer Stichprobe

Vereinfacht ausgedrückt sollen anhand dieser Arbeit textuelle Daten mit potenziellen Raumbezug für Dritte sichtbar gemacht werden. In diesem Zusammenhang wird in der Wissenschaft zwischen drei verschiedenen Methoden zum Verständnisaufbau unterschieden. (Lacity & Janson 1994)

- Positivistisch

  Das Verständnis entsteht durch die Identifikation von nicht-zufälligen Variationen im Text. Es impliziert, dass die Häufigkeit ein wichtiger Indikator ist. Je öfter ein Phänomen auftritt, desto unwahrscheinlicher ist der Zufall. Der Text muss sich dabei auf eine objektive Realität beziehen. Das führt vor allem dann zu Problemen, wenn sprachliche Unsicherheiten vorkommen. Der Satz, „Er ging über die Brücke“ kann auf zwei Arten verstanden werden, da die „Brücke“ zwei Interpretationen zulässt. Es ist möglicherweise ein Bauwerk oder ein Zahnersatz.

- Sprachlich

  Bei dieser Methode wird der Zusammenhang zwischen dem Text und der herrschenden Realität. Ein wichtiger Baustein hierfür ist die Chomsky Grammatik, die von syntaktischen und semantischen Regeln ausgeht. (siehe Kapitel 2.2) Das Verständnis der Realität wird durch das Lernen der sprachlichen Strukturen ermöglicht.

- Interpretativ

  Hierbei wird versucht zu verstehen, wie die Kultur und die Erfahrung die Interpretation beeinflussen. Eine Annahme ist, dass der Forscher oder die Forscherin den Text nicht objektiv analysieren kann. Für die Analyse von historischen Dokumenten gilt, dass sich der Interpret mit dem Autor, der damaligen Kultur und der zeitlichen Einordnung auseinandersetzt.

Der erste Ansatz aus obiger Aufzählung ist den quantitativen Methoden zuzuordnen. Beispiele sind die Sentiment-Analysen von großen Datenmengen oder das Geoparsing von Tweets. Die anderen beiden sind qualitative Methoden, wobei beim interpretativen Ansatz der Forscher oder die Forscherin als Insider fungiert. (Lacity & Janson 1994) Für diese Arbeit wird der sprachliche Ansatz angewandt, denn es gilt einen Zusammenhang zwischen dem vom Gastgeber oder dem Gast Geschriebenen und den Distanzen in der Realität herzustellen. Dies geschieht obwohl Zahlen in Form von Entfernungsangaben untersucht werden und die Herangehensweise somit mehr der quantitativen Methode zuzuordnen ist. Doch nach Hammarberg, Kirkman & de Lacey 2016 sind quantitative Methoden nur dann einzusetzen, wenn die untersuchten Variablen alleine betrachtet werden können. Das ist hier nicht der Fall, denn die Entfernung im Text stellt eine Beziehung zwischen den POIs dar. Nach der Definition der drei Autoren wäre diese Arbeit der „Analyse von Texten und Dokumenten“ einzuordnen, womit über verteiltes Wissen

gelernt wird. Bei der qualitativen Textanalyse ist weniger eine hohe Anzahl an untersuchten Dokumenten wichtig, sondern mehr eine Auswahl an qualitativ hochwertigen. Es sind jene Texte auszuwählen, die Vollständigkeit (dem Thema entsprechend), Selektivität und Balance oder Imbalance (Auswahl von Dokumenten mit hohen und niedrigen Detailierungsgrad) aufweisen. Daraus ergibt sich, dass keine Angabe zur Mindestanzahl an Texten gemacht werden kann. Es ist möglich, dass bereits eine geringe Anzahl, einen effektiven Nutzen zur Erreichung des Forschungsziels beitragen kann. Genauso kann dadurch auch schon eine Sättigung herbeigeführt werden. Das bedeutet, dass trotz des Hinzufügens von neuen Daten, nicht zwingend neue Erkenntnis zu gewinnen sind. (Bowen 2009)

Für eine erste Annäherung wurden alle dreizehn gefundenen Unterkünfte, die in Tabelle 8 aufgelistet sind, herangezogen. Diese Airbnbs haben insgesamt 2.338 Bewertungen, die es zu analysieren gilt. Außerdem hat jeder Gastgeber die Möglichkeit die räumliche Umgebung zu beschreiben. Falls eine solche vorhanden ist, dann fließt auch diese in die Auswertung mit ein. Im Laufe der praktischen Anwendung wird sich zeigen, wie viele Prozent davon der Thematik entsprechen, wie selektiv die Vorgehensweise ist und ob eine Balance oder Imbalance zwischen den Daten herrscht. Besonders der letzte Punkt wird Aufschluss darüber geben, ob die Angabe von räumlichen Informationen wichtiger wird, je weiter entfernt sich die Unterkunft von POIs befindet.

4.2.2 Datenkollektion für die Untersuchung

Die dreizehn ausgewählten Unterkünfte sind in Tabelle 8 aufgelistet. Diese sind nach der Anzahl an Bewertungen absteigend sortiert. In der Tabelle enthalten sind der offizielle Name des Airbnb (Stand: 27.04.2019), der Wiener Gemeindebezirk, in dem sich die Unterkunft befindet und die zur Verfügung stehenden Bewertungen je Airbnb. Eine weitere Spalte stellt die eigens zugewiesene Identifikationsnummer dar. Ein Verweis auf eine bestimmte Unterkunfts-ID bezieht sich in Zukunft in dieser Arbeit immer auf diese Nummer. Die Namen der Gaststätten sind veränderbar, weshalb es wichtig ist, den obigen Zeitstempel zu beachten. Dies stellt die Datenbasis für alle folgenden Analysen dar.

Offizieller Name des Airbnbs	Bezirksname	Anz. Bewertungen	ID
Central, spacious, quiet, bright	Rudolfsheim-Fünfhaus	433	1
VIENNA CENTER – BELVEDERE 22	Landstraße	339	2
Cozy Studio in the City Centre	Leopoldstadt	310	3
Hotspot Museum, first district	Mariahilf	292	4
SUITE - 10 hotspots in 10 minutes	Margareten	211	5
Feel at home with private! bathroom	Ottakring	155	6
Luxury Stephansplatz Suite, directly in the Center	Innere Stadt	153	7
Cosy and Bright Studio near Naschmarkt and Centre	Margareten	112	8
Colorful room for cat lovers	Wieden	96	9
Brand new apartment 5 min to the city center	Favoriten	90	10
Steiner Residences Apt. Picasso (13)	Leopoldstadt	82	11
VIENNA PAINTINGS Künstlerappartment	Währing	40	12
Studio 1140 • By Schönbrunn & Near City Centre	Penzing	25	13

Tabelle 8 Liste der Unterkünfte inklusive der Anzahl an Bewertungen

4.3 Verwendete Technologien

Im Fachbereich der (Geo-) Datenanalyse gibt es eine Vielzahl an unterschiedlicher Software und Programmiersprachen, die den gewünschten Zweck erfüllen können. Dieses Kapitel soll erläutern, welche Technologien im Rahmen dieser Arbeit verwendet werden und warum.

4.3.1 R

R ist aus der statistischen Programmiersprache S entstanden und hat die Notwendigkeit der Definition von globalen Variablen im Code entfernt. Mit diesem Paradigma sollte die Sprache intuitiver zu benutzen sein. Der Code wird außerdem nicht kompiliert, sondern direkt ausgeführt. (Becker) Als Teil des GNU-Projektes (General Public License) ist R frei verfügbar und zu verwenden. Über ein umfassendes R Archiv-Netzwerk (engl. Comprehensive R Archive Network, CRAN) ist es mit Stand 22.06.2019 möglich, 14.445 geprüfte Pakete herunterzuladen, welche die Funktionalitäten von R erweitern. (CRAN 2019) Für die Programmierung selbst wird auf die grafische Oberfläche von RStudio

zurückgegriffen. Diese ermöglicht einen Überblick über die deklarierten Variablen, über die geladenen Paketen und es zeigt Syntaxhervorhebungen an.

Wenn im Bereich der Datenanalyse von R gesprochen wird, so muss auch immer auf die Programmiersprache Python eingegangen werden. Python beschränkt sich nicht nur auf die Analyse an sich, sondern wird als Softwareentwicklungs-Werkzeug angesehen. In der Tabelle 9 sind die jeweiligen Unterschiede und Gemeinsamkeiten der beiden Sprachen angegeben. Wobei klar ersichtlich ist, dass es beinahe nur Überschneidungen gibt. R und Python speichern und rechnen Daten im RAM (Random-Access Memory), bei einem Fehler wird immer die genaue Zeilennummer angegeben und es gibt auch Bibliotheken und Methoden, um fehlende Werte auszubessern. Der markanteste Unterschied ist bei der Anzahl an verfügbaren Paketen zu erkennen. Beide Sprachen bauen auf einer breiten Online-Gemeinschaft, welche Bibliotheken (weiter-) entwickelt, auf. (Brittain et al. 2018) Die stärkere Beliebtheit von Python ist bei den Fragen auf der Internetplattform *Stack Overflow* zu sehen. Bei einem Vergleich der Entwicklung der Klicks über alle Programmiersprachen zwischen 2012 und 2019 bei Fragen mit einem jeweiligen *Tag*, hat R eine Entwicklung von ~0,6% auf ~2,8% und Python von ~3,5% auf ~12,2%. (StackOverflow 2019) Der Unterschied lässt sich teilweise dadurch erklären, da Python als übergeordnetes Ziel den produktiven Einsatz hat, während sich R auf die Erzeugung von Datenanalyse-Modelle fokussiert. (Brittain et al. 2018) Die anderen Funktionalitäten der beiden Technologien sind sonst de facto ident, wie Tabelle 9 zeigt. Es ist also dem Anwender und dessen Zielsetzung überlassen ob R oder Python verwendet wird. Das Modell, das in dieser Arbeit Anwendung findet, ist vorwiegend auf den akademischen Bereich beschränkt und soll in einem ersten Schritt generell die praktische Möglichkeit feststellen. Für diese Anforderungen ist R als ausreichend anzusehen.

Attribut	**Python**	**R**
Verfügbare Pakete	184.632	14.445
Daten-Handhabung	RAM	RAM
Zeilenbezogener Fehler	Ja	Ja
Nummern und Text	Ja	Ja
Interaktivität & Oberfläche	CLI und IDE	CLI und IDE
komplexe Datenstrukturen	Ja	Ja
Fehlende Werte	Ja	Ja
Lineare Algebra	Ja	Ja
Grafiken	Ja	Ja
Quelle: Brittain et al. 2018		

Tabelle 9 Gegenüberstellung von R und Python

4.3.2 GIS-Software

Im Rahmen dieser Arbeit werden zwei GIS-Software-Systeme verwendet, denen jeweils ein unterschiedliches Entstehungsparadigma zugrunde liegt.

ESRI ArcGIS

Bei ArcGIS handelt es sich um eine bekannte GIS-Software von ESRI (Environment System Research Institute). ESRI ist der größte Hersteller von proprietärerer GIS-Software weltweit. Die Firma befindet sich in Redlands, Kalifornien und wurde 1969 gegründet. Die Software ArcView 1 war 1992 das erste leistbare benutzerfreundliche Desktop-GIS auf dem Markt. Aus ArcView ist über die Jahre ArcMap entstanden, jene Software, die auch in dieser Arbeit Verwendung findet. (Peters 2015) Die Produktpalette von ArcGIS ist „gut dokumentiert, bietet mannigfaltige Analysetechniken und stellt ein robustes“ GIS dar. (Kotzbek 2016) ArcGIS kann mit anderen Technologien, wie Python oder MATLAB verbunden werden, um spezielle Methoden zu implementieren. (Maogui & Jinfeng 2009)

QGIS

Im Gegensatz zu ArcGIS ist QGIS eine freiverfügbare GIS-Software. Die erste stabile Version 1.0 wurde im Jahr 2009 veröffentlicht. (aktuellste stabile Version: 3.4, Stand: Juni 2019) Mit Hilfe des *Plugin Builder* kann jeder Entwickler eine eigene Erweiterung programmieren, einsetzen und diese auch veröffentlichen. Es gibt ein QGIS Kern-Entwicklerteam, das aus 112 aktiven Mitgliedern besteht. Die Installation von solchen Erweiterungen ist ein großer monetärer Vorteil gegenüber ArcGIS. Denn für die Nutzung von diversen Erweiterungen, wie zum Beispiel dem Network-Analyst, ist zu bezahlen bzw. muss auf eine höhere Version umgestiegen werden. Wie R ist auch QGIS unter der GNU-Lizenz verfügbar und ist deshalb uneingeschränkt nutzbar. (Nicolas & Frédéric 2018)

Es gibt drei wesentliche praktische Schritte, die die ausgewählte GIS-Software erledigen muss: (1) erstellen eines Routing-Netzwerks auf Basis der GIP, (2) erzeugen von Puffer rund um die POIs auf Basis der Distanzangabe, die in einer bestimmten Spalte enthalten ist und (3) darstellen der Ergebnisse auf einer bzw. mehreren Karten deren Layout immer ident ist. Für das Routing-Netzwerk gibt es von ArcGIS den sogenannten *Network-Analyst* und von QGIS das *Road Graph* Plugin. Beide Methoden erfüllen die Anforderung, aber in diesem Fall wird auf ArcGIS zurückgegriffen, da es doch essenziell ist, dass das erstellte Netzwerk eine hohe Güte aufweist und der Network-Analyst auch schon in anderen akademischen Arbeiten eingesetzt wurde. Es ist nur mit Hilfe von ArcGIS möglich einen Puffer mit einer bestimmten Distanz auf Basis einer Spalte zu erzeugen. In der neuesten stabilen QGIS-Version 3.4 gibt es eine solche Funktionalität nicht. Die Entscheidung, ob die Karten in ArcGIS oder in QGIS erstellt werden sollen, erfolgt auf der Überlegung wie einfach es ist, mehrere Karten mit dem identen Design zu erstellen, ohne dabei ständig ins Layout eingreifen zu müssen. Dazu hat Friedrich 2014 in ihrer Arbeit gezeigt, dass vor allem bei der Kartensyntax und der Kartenkonstruktion so gut wie kein Unterschied zwischen den beiden Systemen besteht. Es könnten also grundsätzlich sowohl ArcGIS als

auch QGIS für die Lösung der Kartendarstellung mitwirken – sie sind in dieser Frage als gleichwertig anzusehen. Im Fall dieser Arbeit fiel die Entscheidung aufgrund der Freiverfügbarkeit auf QGIS.

4.4 Praktische Aufarbeitung

Diese Kapitel widmet sich der Analyse, der Bewertungen und der Beschreibungen von Airbnbs. Es soll die Herausforderungen und deren Lösungen im Zusammenhang mit der Extraktion von räumlichen Daten aus textueller Information zeigen. Zum Schluss des Abschnitts wird erläutert, wie die Informationen, die weiters in Kapitel 5 einfließen, zustande gekommen sind.

Aus inhaltlicher Sicht und aus Gründen der Verarbeitungsgeschwindigkeit werden aus den vorhandenen 11.276 Airbnbs in Wien, nur jene dreizehn selektiert, die in Tabelle 8 gelistet sind. Somit ist garantiert, dass auch nur die dazugehörigen Bewertungen verarbeitet werden. Die POIs werden in diesem Abschnitt nicht verändert.

Nach dem Einlesen der Bewertungen und Beschreibungen, erfolgt die Tokenisierung der Sätze. Das führt dazu, dass ein Satz in dessen einzelnen Wörter zerlegt wird. Ein Satz, der beispielsweise aus zehn Wörtern besteht, wird in diesem Schritt in zehn separate Zeilen aufgeteilt. Die Punktationen fallen bei dem Schritt heraus und können nicht mehr hergestellt werden. Zusätzlich zu den Wörtern bleiben allerdings auch Zahlen bestehen. Diese sind für die Extraktion in Kapitel 4.4.2 essenziell.

Die gesamte Vorgehensweise der Aufbereitung mittels R erfolgte aus eigenen Überlegungen. Bei der Verwendung von Code-Fragmenten aus fremden Quellen wurde es im Skript angegeben. Die genutzten Pakete, deren allgemeine Beschreibung und die wichtigsten Methoden für diese Arbeit sind in Tabelle 10 aufgelistet.

4.4.1 Wichtige Paradigmen in R

Eine der wichtigsten Datenstrukturen in R ist der *Dataframe*. Es handelt sich dabei um eine Liste an Vektoren, die die gleiche Länge aufweisen. Deshalb wird beim Verschmelzungsversuch von zwei oder mehrere Vektoren, die unterschiedliche Längen aufweisen, ein Fehler zurückgegeben. Bei erfolgreicher Erstellung besteht ein Dataframe aus Spaltennamen (Variablen) und Zellwerten. Eine gesamte Zeile an Zellwerten wird auch Beobachtung genannt. Eine Information, die nur in seltenen Fällen verwendet wird, ist der Zeilenname. Jeder erstellte Dataframe in R hat solche Zeilennamen. Diese sind Zeichenketten, welche genausooft vorkommen, wie es Zeilen gibt – für jede Zeile gibt es demnach einen eindeutigen Namen. Außerdem kann es keine fehlenden Werte geben.

Paket	Methoden	Beschreibung
dplyr	`filter()`	Je nach Werten können bestimmte Beobachtungen selektiert werden.
	`mutate()`	Erzeugt eine neue Variable auf Basis von Funktionen und existierenden Variablen
	`group_by()`	Teilt einen Dataframe in bestimmte Gruppen an Zeilen
	`select()`	Selektiert nur bestimmte Variablen
tidytext	`unnest_tokens()`	Eine Zelle, die mehrere Wörter enthält, wird in genauso viele Spalten und einzelne Wörter zerlegt. So wird aus einem Satz mit 5 Wörtern, eine neue Spalte mit 5 Zeilen erstellt.
stringr	`str_length()`	Wird eingesetzt, um mit Hilfe der Funktion *substr()*, bestimmte Buchstaben eines Wortes in einem automatisierten Prozess zu extrahieren.
readr	`parse_number()`	Filtert alle Zahlen, die als Wörter tokenisiert wurden. Das ist wichtig für die Suche nach Distanzangaben, die in Zahlen ausgedrückt sind.

Tabelle 10 Eingesetzte Pakete und deren wichtigsten Funktionen

Das *Tidyverse* beinhaltet eine Fülle an Paketen für R, die ein „harmonisches“ Arbeiten durch einheitliche Datenrepräsentation und Schnittstellen sicherstellt. (Wickham 2019a) Die *dplyr*-Bibliothek aus dem Tidyverse dient zur Datenmanipulation. Das zentrale Element ist die *Pipe*, die mit „%>%“ innerhalb des Codes dargestellt wird. Eine Eingabe, die vor der Pipe aufgelistet ist, dient als Inhalt für den nachfolgenden Schritt. Beim Lesen des Codes wird nicht Pipe gesagt, sondern „dann“. Wenn das Ergebnis nach dem Pipe-Prozess in eine Variable fließen soll, dann ist der Befehl „->“ notwendig. (Wickham 2019b) Ein beispielhafter Datenfluss ist in Abbildung 16 dargestellt.

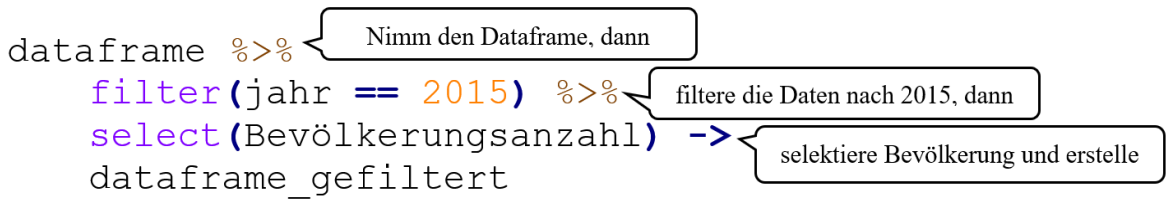


Abbildung 16 Lesen des Codes unter der Verwendung von Pipes

4.4.2 Extraktion räumlich relevanter Information

Bei der Tokenisierung von Sätzen werden nur jene Wörter aufgeteilt, die durch eine Leertaste voneinander getrennt sind. Das hat zur Folge, dass bei Distanzangaben, die numerische Entfernungsangabe und dessen Einheit ohne Leertaste angegeben sind. Beispiele hierfür sind: *200m* oder *10min*. Diese werden vom System auch nicht als numerischer Wert erkannt, weshalb ein Abgleich über den letzten Buchstaben des Tokens erfolgen muss. Ein möglicher relevanter, räumlicher Token wird dann markiert, wenn ein

„*m*“ (Endung für Meter, m) oder „*t*“ (Endung für Feet, ft) in der Zelle vorkommt. Für zeitliche Angaben wurde „n“ (Endung für Minuten, min) für den Abgleich festgelegt.

Wenn ein Abstand zwischen Zahl und Einheit vorhanden ist, dann kann über ganze Wörter abgeglichen werden. Für einen solchen Fall ist ein Lexikon entstanden. Unter einem Lexikon wird in dieser Arbeit eine Tabelle verstanden, die Ausprägungen von räumlichen oder zeitlichen Distanzangaben enthält. (Tabelle 11) Die Lexika wurden selbst erstellt und sollen helfen, die relevanten Sätze zu finden. Wenn ein im Lexikon vorkommendes Wort getroffen ist, dann wird es markiert. Die dahinterstehenden Annahmen sind, dass zeitliche Angaben in einem städtischen Maßstab nicht in Stunden oder Sekunden erfolgen und durch die Berücksichtigung von englischsprachigen Texten auch die angloamerikanischen Längeneinheiten vorkommen können.

Zusammenfassend sind zwei Annahmen, wie Distanzangaben in schriftlicher Form vorhanden sein können, getroffen worden. Entweder durch *XXX‿[räumlicher/zeitlicher Begriff aus dem Lexikon oder bestimmte Endung des Tokens]* oder durch *XXX[räumlicher/zeitlicher Begriff aus dem Lexikon oder bestimmte Endung des Tokens]*.

Distanzangaben	
räumlich	*zeitlich*
m	Minuten
Meter	minutes
mtr	min
mts	mins
meter	
feet	
ft	
meters	
Meters	
mi	
Mile	
miles	
mile	

Tabelle 11 Lexika zu Distanzangaben

Nach der erfolgten Markierung der relevanten Wörter bzw. Wortpaare liegt noch immer eine große Tabelle vor. Das bedeutet, dass die Information bis jetzt nicht verringert wurde. Dieser Umstand sorgt für Performanceeinbußen, da theoretisch wieder jeder Token analysiert werden muss. Die Lösung stellt eine Verkleinerung des Korpus, um die irrelevante Information dar. Bei einer Verkleinerung des Korpus ist es wichtig, dass kein Informationsverlust erfolgt. Diese Bedingung wird dadurch erfüllt, dass die vorherigen und nachfolgenden zwölf Wörter rund um die markierten Wortfolgen selektiert werden. Die Anzahl an ‚zentralen‘ Wortfolgen für die Verkleinerung des Korpus ist in Tabelle 12 dargestellt. Es sind auch Wortfolgen mithineingeflossen, die keine Distanzangabe an erster Position in Zahlenformat besitzen, sondern zum Beispiel „einige“ oder „wenige“. Wieder bezugnehmend zum Inhalt von Tabelle 12 ist klar ersichtlich, dass wohl der meiste, zu

verarbeitende Inhalt aus den Bewertungen stammen wird. Obwohl dreizehn Unterkünfte ein Teil der Analyse sind, haben davon nur acht eine Beschreibung zur öffentlichen Anbindung eingefügt.

Bewertungen					
Anzahl	enthaltene Wörter	gefundene Wörter			
		räumlich ohne Leerzeichen	zeitlich ohne Leerzeichen	räumlich mit Leerzeichen	zeitlich mit Leerzeichen
2.338	114.303	35	39	36	287

Beschreibungen					
Anzahl	enthaltene Wörter	gefundene Wörter			
		räumlich ohne Leerzeichen	zeitlich ohne Leerzeichen	räumlich mit Leerzeichen	zeitlich mit Leerzeichen
8	573	2	0	0	8

Tabelle 12 potenziell relevante räumliche und zeitliche Wortfolgen

Für das Finden der insgesamt 24 umgebenden Wörter war eine eigene Funktion in R notwendig. In einem ersten Schritt musste ein Schlüsselwort gesucht und gefunden werden um darauffolgend die vorherigen und nachherigen zwölf Wörter pro Schlüsselwort aus den jeweiligen Zeilen zu extrahieren.

```
# Definition der Funktion
# Quelle: https://stackoverflow.com/questions/13155609/returning-above-and-below-rows-of-specific-rows-in-r-dataframe
(aufgerufen am: 11.03.2019)
getMyRows <- function(data, matches, range) {
  rowMatches = lapply(unlist(lapply(matches, function(x) grep(x,
               rownames(data)))), function(y) y + range)
  rowMatches = lapply(rowMatches, function(x) x[x > 0 & x <=
               nrow(data)])
  lapply(rowMatches, function(x) data[x, ])
}

# beispielhafter Aufruf der Funktion getMyRows()
spatial_temp <- getMyRows(listings_spatial_temp, c("Yes"), -12:12)
```

Der obigen R-Funktion `getMyRows()` werden drei Parameter übergeben. Das ist zum einen der Dataframe, in dem jede Zeile markiert wurde, in der ein Wort aus dem Lexikon oder ein Hinweis auf eine räumliche oder zeitliche Distanzangabe vorhanden ist. Des Weiteren wird das zu suchende Wort, das Schlüsselwort und der umgebende Bereich des gefunden Wortes übergeben – in diesem Fall zwölf Wörter davor und danach. In der ersten Zuweisung wird mit der inneren Funktion `lapply()` das zu treffende Schlüsselwort an eine anonyme Funktion übergeben. Mit `grep()` wird mittels regulären Ausdrücken untersucht, ob das Schlüsselwort im Zeilennamen des übergebenen Dataframes vorhanden

ist. Hier kommen die oben erwähnten Zeilennamen zum Einsatz. Mit deren Vorhandensein ist es bei dieser Vorgehensweise möglich die gesamten Zeilen zu selektieren. Die daraus entstehende Liste wird wiederum mittels `unlist()` in einen Vektor verwandelt, um später wieder als Eingabe in einem `lapply()` für eine anonyme Funktion zu dienen. Dies macht Sinn, weil ein `lapply()` nicht auf ein `lapply()` folgen kann. In dem letzten Schritt der ersten Zuweisung werden der gefundenen Position die vorherigen und nachherigen zwölf Zeilennummern übergeben. Um Fehlermeldungen in der Ausführung der Funktion zu vermeiden erfolgt eine Kontrolle, um sicherzustellen, dass sich keine Position außerhalb des übergebenen Dataframes befindet. Im letzten Schritt werden alle Zeilen selektiert, welche die bestimmten Zeilennummern enthalten. Bei Aufruf der Funktion enthält somit der daraus entstehende Dataframe alle relevanten Zeilennummern bzw. Zeilennamen, welche für die spätere Analysen von Relevanz sind. Der Nachteil dieser Vorgehensweise ist, dass auch Wörter selektiert werden, die eigentlich nicht zu dem Satz gehören, in dem das räumlich relevante Wort vorkommt. Dies muss vor dem Export beachtet und über die IDs der einzelnen Sätze bereinigt werden.

Die in Tabelle 12 angeführten Zahlen, sind all jene, die eine potenziell zeitliche oder räumliche Distanzangabe beinhalten. In einem nächsten Schritt wurden auch noch all jene herausgefiltert, die einen vorrangreihten Zahlenwert beinhalten. Unterm Strich befinden sich dann insgesamt 293 Sätze im Schritt der Kodierung. Es folgt eine Einteilung der Sätze nach vier Kategorien.

- 0: Der Satz weist keine räumliche oder zeitliche Relevanz auf. Ein Grund für die Fehleinschätzung bei der vorherigen Filterung ist beispielsweise die Angabe von Uhrzeiten im Zwölf-Stunden-System.
- 1: Im Satz befindet sich eine Distanzangabe, doch es unklar auf welchen POI sich diese bezieht.
- 2: Es ist eindeutig herauszulesen, auf welchen POI sich die Distanzangabe im Satz bezieht. Die Sätze mit diesem Code können später in der Analyse im GIS verwendet werden.
- 3: Der Satz ist nicht in deutscher oder englischer Sprache gehalten. Es kann auch sein, dass ein Fehler in der Kodierung vorliegt.

Der angewandte Korpus enthält außerdem 9.107 Sätze. Diese Zahl wurde in Tabelle 12 nicht angeführt, da die Sätze als eigenständiges Element für diese Arbeit nicht so wichtig waren, wie die einzelnen Wörter. Es gibt keinen Vergleichswert aus anderen Studien, welche das Vorhandensein von quantitativen räumlichen Ausdrücken untersuchte. Der in Kapitel 2.4.4 angegeben Anteil von 2% an räumlichen Ausdrücken im sozial Netzwerk Twitter bezog alle Sätze mit ein, die aus einem Lokatum, einem Referenzobjekt und einer Präposition (räumliche Beziehung) bestanden. Hier gilt allerdings eine doch selektivere Auswahl, da nur quantitative Beziehungen von Relevanz waren. Von den 9.107 Sätzen konnten 59 Sätze mit einer 2 kodiert werden. Das entspricht einem Anteil von 0,65%.

Bei der NER wird üblicherweise die Genauigkeit und der Ausfall als Qualitätsmaß herangezogen. Wichtig dafür sind die Werte für wahrlich positiv (korrekt erkannte Tripletten/Entitäten), falsch positiv (falsch erkannte Tripletten/Entitäten) und falsch negativ (es handelt sich eigentlich um relevante Information, doch vom System wurde es nicht als solche erkannt). (Zhang 2013: 31)

$$Genauigkeit = \frac{|wahrlich\ positiv|}{|wahrlich\ positiv| + |falsch\ positiv|} \tag{3}$$

$$Ausfall = \frac{|wahrlich\ positiv|}{|wahrlich\ positiv| + |falsch\ negativ|} \tag{4}$$

Die Genauigkeit beschreibt demnach wie gut ein System zur NER die Entitäten richtig vorhersagt und der Ausfall gibt an, in welchem Ausmaß die räumlichen Objekte überhaupt gefunden werden. (Zhang 2013: 31)

Da es sich hierbei um einen teilautomatisierten Prozess handelt, war es nicht möglich die Anzahl an falsch negativen Tripletten zu ermitteln. Dazu waren die Annahmen, die getroffen wurden, zu restriktiv. Nur somit konnten überhaupt Tripletten mit metrischer Information gefunden werden. Das Festlegen von falsch negativ erkannten Objekten ist wesentlich einfacher, wenn der Kontext nicht von Bedeutung ist. Daraus schlussfolgernd kann für diese Arbeit nur die Genauigkeit berechnet werden. Dabei sind die wahrlich positiven, alle 59 detektierten, die einen Beitrag zur räumlichen Analyse leisten könnten. Die falsch positiven Markierungen ergeben sich aus den insgesamt 293 Sätzen abzüglich der 59, also 234. Daraus ergibt sich eine Genauigkeit von ~20%. Diese berücksichtigt alle Hypothesen aus Kapitel 4.4.2 und eine händische Klassifikation.

In Anhang III ist sind alle 293 Sätze aufgelistet, die als „potenziell räumlich“ markiert wurden. Es sind zwar alle Sätze gefunden worden, die eine räumliche oder zeitliche Entfernung in quantitativer Form enthalten, allerdings sind auch Redundanzen in den minimierten Datenbestand gelangt. Dieser Umstand ergab sich dadurch, dass die zwölf Wörter vor und nach einer detektierten räumlichen Relation selektiert wurden. Es erfolgte keine Detektion ob es noch der gleiche Satz war, da die Annahme bestand, dass das Lokatum oder das Referenzobjekt sich auch in einem vorherigen Satz befinden könnte. Insgesamt sind 36 Zeilen als redundant einzustufen, da sie keine sinnvolle räumliche oder zeitliche Information beinhalten oder falsch kodiert sind. Außerdem beschreiben nicht alle Gastgeber die Unterkunft mit einer quantitativen Beschreibung zur Umgebung. Von den 13 untersuchten Airbnbs haben 7 Hosts die zurückzulegenden Wege aufgelistet. Bei der Unterkunft mit der ID 3 wurde eine mehrfach ideale Auskunft gegeben. Dabei schrieb der Host folgende Beschreibung: „*…from this u1 station nestroyplatz is a 2 minute walk away, u4 schwedenplatz is 8 minutes away and u2 praterstern is 9 minutes away…*“. In einem einzigen Satz sind demnach drei unterschiedliche und eindeutig benannte Referenzen enthalten. Diese relative lange Tabelle ist deshalb angehängt, damit die Nachvollziehbarkeit der Kodierung gegeben ist, denn diese wurde händisch und nicht automatisiert vom Computer bzw. R durchgeführt. Das hat mehrere Gründe:

- Zum einen wird dadurch eine höhere Qualität des späteren Analyseergebnisses erreicht. Das Hauptziel der Arbeit ist es festzustellen, ob überhaupt die Möglichkeit besteht aus nutzergenerierten Daten den Standort festzustellen. Für die Auswertung ist wichtig, dass so viele potenziell relevanten Daten wie möglich gefunden werden. Durch die händische Kodierung konnten 59 von 293 Sätzen als vollkommen bedeutsam markiert werden. Bei einer automatisierten Bewertung wäre diese Anzahl deutlich niedriger. Die zwei Fehlerquellen, die dazu geführt hätten, sind aus den folgenden beiden Punkten herauszulesen.

- Was für den Menschen eindeutig herauszulesen ist, ist für die Maschine eine Herausforderung. Vor allem die Identifikation der jeweiligen Referenzierung kann zu einer unsicheren Interpretation führen. Die beiden folgenden Beispielsätze aus dem Korpus sollen helfen das Problem zu verstehen:

    - *neighborohood it is well located just few steps from transport which takes 10 minutes to get karlsplatz.*

  In diesem Satz ist für die Maschine unklar, dass der Gast zuerst in die Straßenbahn oder die U-Bahn einsteigt und mit dieser zehn Minuten benötigt. Die Unterkunft ist nicht zehn Minuten vom Karlsplatz entfernt.

    - *only 5min from the metro and 12min walk from the airport train link station which is always nice.*

  Dieser Satz zeigt eine ähnliche Komplexität wie oben. Es ist unsicher worauf sich die Distanzangaben im Einzelnen beziehen.

- Neben der Distanzangabe wurden auch die POIs händisch herausgelesen. Es hat sich gezeigt, dass ein und derselbe Ort unterschiedliche Ausprägungen annehmen kann. Ein daran angepasstes Gazetteer hätte einen Mehraufwand erzeugt, der sich nicht lohnt, da beim Hinzukommen neuer Bewertungen, dies erneut händisch angepasst hätte werden müssen. Das folgende Beispiel zeigt wie viele Ausprägungen ein einziger POI, verfasst von unterschiedlichen Nutzern verursachen kann. Im folgenden Beispiel wird der Wiener Hauptbahnhof als der am häufigsten vorkommende PIO in all seinen geschriebenen Ausprägungen dargestellt.

    - *wien hauptbahnhof* (Angabe vom Gastgeber), *main train station, main rail station, central station, vienna central train station, wien station, central station wienhauptbahnhof* (sic!), *vienna's hauptbahnhof, hbf, central railway station, haubtbahnhof* (sic!)

  Unabhängig von der Groß- und Kleinschreibung betrachtet, wird der Wiener Hauptbahnhof in dieser beispielhaften Betrachtung auf elf unterschiedliche Arten und Weisen geschrieben. Davon ist die Schreibweise zweimal falsch und einmal eine Abkürzung.

Im ersten Punkt der obigen Auflistung ist bereits auf die Quantitäten eingegangen worden. In Abbildung 17 wird dies noch granularer dargestellt. Sie macht deutlich, dass die

Unterkünfte mit der ID 11, 12 und 13 bereits aus allen nachfolgenden Analysen ausscheiden, da hier keine oder nur eine räumlich relevante Information vorliegt. Wenn maximal eine räumliche Angabe vorhanden ist, dann können keine Analysen im GIS durchgeführt werden. Für die Bildung von Überschneidungsbereichen werden zumindest zwei Distanzangaben, ausgehend von zwei verschiedenen PIOs benötigt.

Außerdem weisen viele Sätze keine räumliche oder zeitliche Relevanz auf (31,3%). In diesen Sätzen sind Angaben zur Uhrzeit, zum Beispiel wann eine Person angekommen oder abgeholt wurde oder Informationen zu einer bestimmten Fahrdauer mit einem der öffentlichen Verkehrsmittel. Die Uhrzeit wird auch deshalb falsch erkannt, da durch den Abgleich über den letzten Buchstaben bei potenziellen Distanzangaben „m“, die Uhrzeiten im 12-Stunden-Format (zum Beispiel: 12.00am oder 10.30pm) als relevant erkannt wurde. Der Code mit der Nummer eins hatte einen Anteil von 37,8%. Das sind all jene Fälle, bei denen der referenzierende POI nicht aus den Sätzen herauszulesen war. In Abbildung 17 ist auch der Anteil an relevanter räumlicher Information je Unterkunft errechnet.

$$Anteil\ räumlich\ relevanter\ Information\ [\%] = \sum_{i} \frac{Anzahl\ Code\ 2}{Anzahl\ Sätze} \tag{3}$$

Die Unterkunft mit der ID 1 hat zwar die meisten Kommentare, allerdings auch die zweitwenigsten Sätze (bei Ausschluss von ID 11, 12 und 13) mit räumlich oder zeitlich relevanter Information. Ein statistischer Test zur Korrelation sollte feststellen, ob dies zufällig ist oder nicht. Nach der Durchführung des Tests zur Detektion des Korrelationskoeffizienten nach Pearson hat sich gezeigt, dass bei einem Signifikanzniveau von 0,05, die Anzahl an Kommentaren mit dem Vorkommen an räumlich relevanten Sätzen korreliert. Der Korrelationskoeffizient beträgt in diesem Fall +0,6.

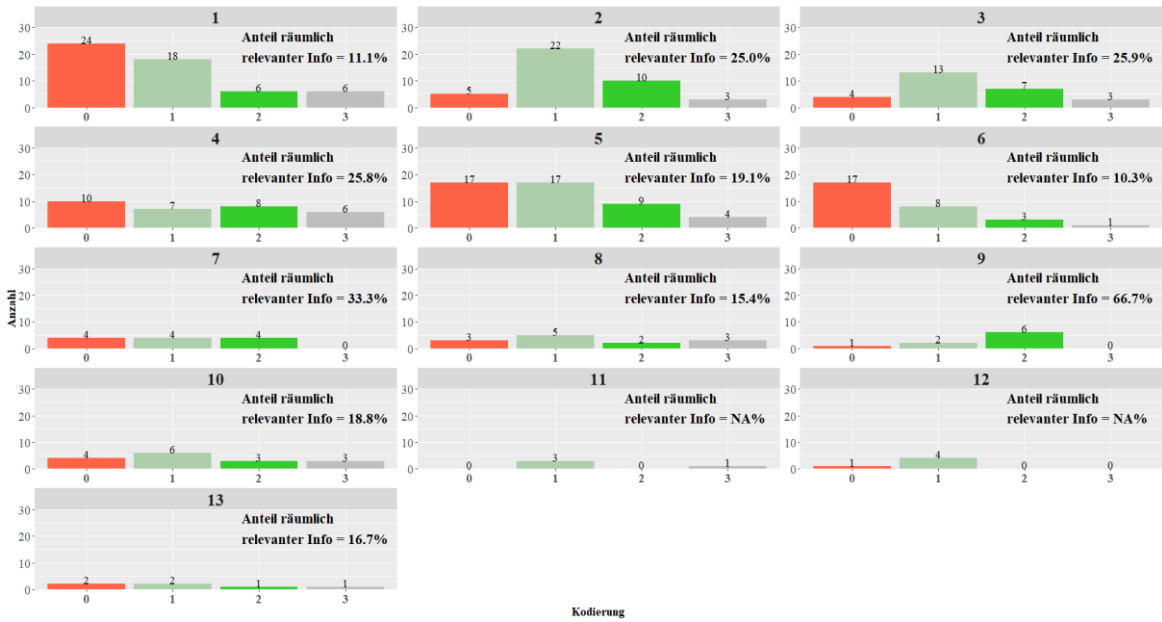


Abbildung 17 Anzahl an Codes je Airbnb

Nur anhand der in Abbildung 17 abgebildeten Werte war es bereits möglich drei Unterkünfte auszuschließen. Eine hohe Quantität bei Code 2 bedeutet allerdings nicht, dass nicht immer nur der idente POI erwähnt wird. Dieser Umstand würde auch eine sinnvolle Auswertung in einem GIS verhindern. Für die Findung eines Standortes bzw. zur Einengung des suchenden Korridors werden mindestens zwei verschiedene POIs benötigt. Eine Zählung ergab, dass nur bei den Unterkünften mit der ID 1, 2, 3, 4, 5 und 7 diese Anforderung erfüllt ist. Alle anderen Airbnbs werden in der folgenden Auswertung nicht mehr berücksichtigt.

Aufgrund der Anzahl an unterschiedlichen POIs haben die Airbnbs mit der ID 2, 4, und 5 das höchste Potential eine sinnvolle Datengrundlage für die Analyse im GIS zu bieten. Diese Tatsache ist darauf zurückzuführen, da bei diesen Airbnbs die Anzahl an eindeutigen POIs am höchsten ist. Bei den Unterkünften mit der ID 1 und 7 stehen nur jeweils zwei verschiedene POIs für die Auswertung zur Verfügung. Am häufigsten wird in den Bewertungen und Beschreibungen auf den Stephansdom und die Wiener Staatsoper referenziert. Der Stephansdom ist das Zentrum der Stadt Wien und liegt direkt in der Innenstadt. Am zweit- und dritthäufigsten wir der Wiener Westbahnhof bzw. der Wiener Hauptbahnhof genannt. Allerdings ist anzumerken, dass der Westbahnhof nur bei einer Unterkunft erwähnt wird – im Gegensatz dazu wird der Stephansdom bei vier von sieben Gaststätten genannt. (Abbildung 18)

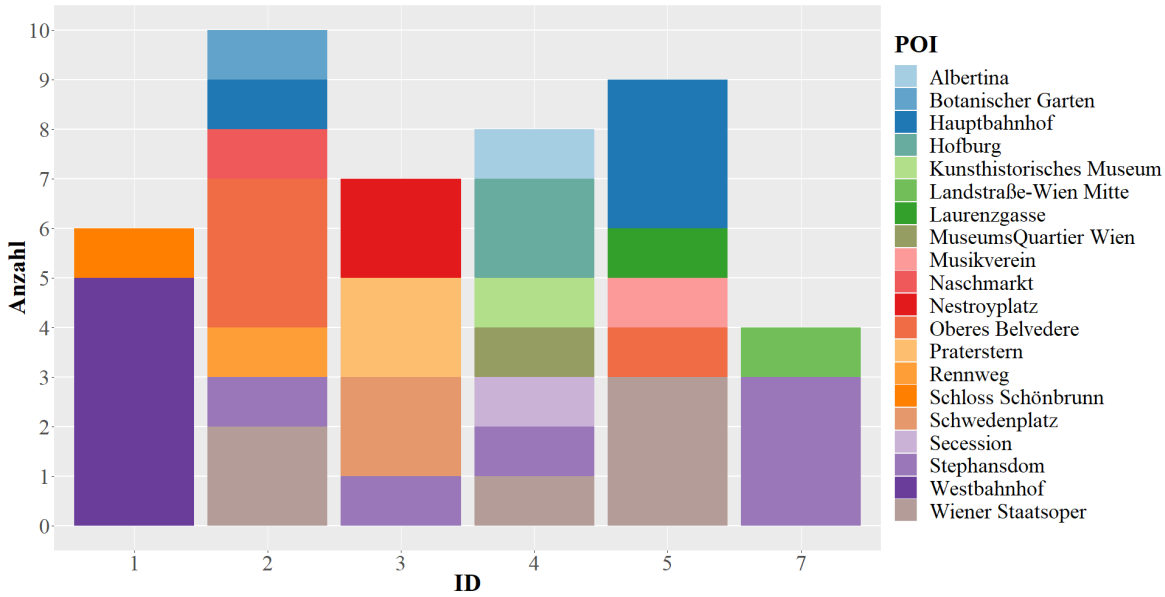


Abbildung 18 Anzahl eindeutiger POIs je Airbnb

Das Gazetteer bilden im wesentlich die POIs der Stadt Wien, die im Kapitel 3.4 aufgelistet sind. Die räumlichen Informationen liegen im Referenzsystem WGS84 vor. Durch einen Abgleich zwischen den Namen der POIs und den Referenzorten aus den Kommentaren und Beschreibungen, war es möglich den Längen- und Breitengrad aus den offiziellen Daten zu übernehmen.

4.4.3 Ethischer Grundgedanke

In dieser Arbeit soll auf Basis der oben angeführten Unterkünfte eine Methodenevaluierung stattfinden. Das Ziel der Arbeit ist nicht das Airbnb bzw. den Vermieter zu veröffentlichen – dies haben die Gastgeber bei den Trainingsdaten bereits selbst getan. Nichtsdestotrotz ist es wichtig auf den Begriff der Ethik im Bereich von Datenanalyse einzugehen, denn damit soll der Verantwortungsgedanke gestärkt werden.

Eine von mehreren Definitionen von Ethik nach Krämer 1992:

> *„'Ethik' liegt dem Wortsinn nach primär immer dann vor, wenn in der Lebensführung sich feste Formen des ‚Ethos', also Grundhaltungen und Verhaltens- oder Handlungsdispositionen abzeichnen, aus denen heraus die einzelnen Handlungen oder Verhaltensweisen vollzogen werden. ... Jedes zielorientierte Handeln oder Sachverhalten, das reguliert verläuft und darin auf konstante Haltungen und Orientierungen verweist, erfüllt schon den Minimalbegriff eines Gegenstandes von Ethik, ja jede einzelne geregelte Handlungsfolge ist potenziell schon ethisch relevant."*

Der Fachbereich des Data-Scientist hat sich in den letzten Jahren stark hin zu großen Datenmengen gewandelt. Neue Technologien, wie arbeiten an unterschiedlichen Standorten, die Vernetzung von Regierung, Industrie und Forschung oder Cloud-Computing haben die Datenmengen rasant steigen lassen. Eine Aufgabe des Data-Scientist besteht nun darin, aus den eventuell unstrukturierten Daten einen Mehrwert zu generieren. Bei dieser Arbeit ist es schwierig die Grenze zu ziehen, wer für die Datenweiterverarbeitung verantwortlich ist. Ein Beispiel hierzu sind die sozialen Medien: Facebook, WhatsApp, Instagram, etc. Der Eintritt in diese Medien ist freiwillig und muss vom Anwender aktiv vollzogen werden. Bei der Anmeldung wird übermittelt, welche Daten den Anbietern nun zur Verfügung stehen könnten. Werden diese AGBs abgelehnt, so kann man nicht beitreten und der Anbieter kann nicht auf meine Daten zugreifen. Im umgekehrten Fall stellt sich die Frage, wieso ein Data Analyst dafür verantwortlich ist, wenn eine Person ihm aus freien Stücken seine persönlichen Daten gibt und er diese auswertet. Es dürfte aber eher um die Absicht der Auswertung an sich gehen. Es wäre heutzutage zum Beispiel ethisch nicht korrekt auszuwerten, ob ein homosexueller Mann mehr weibliche als männliche Freunde auf Facebook hat. Man könnte sich hierbei immer die Frage stellen, ob es in einem alltäglichen Gespräch wäre, die Person mit einer solchen Frage zu konfrontieren. (Leonelli 2016)

Die Ethik ist daher aus der oben beschriebenen Sicht nicht unbedingt ein „Add-on", sondern in der Auswertung allgegenwärtig. Folgend muss einem Data Scientist danach auch bewusst sein, wie er die Daten auswertet und ob er diese auch öffentlich zugänglich macht. Wer eine Aktion setzt, ist für diese auch immer verantwortlich. (Leonelli 2016)

5 Implementierung in ein GIS

Dieses Kapitel zeigt die notwendigen Vorarbeiten auf, die zu tätigen sind, bevor eine Analyse in einem GIS durchgeführt werden kann. Außerdem folgt eine Erklärung, wie eine multiple Distanzangabe je POI und Airbnb gehandhabt wird. Den Abschluss bildet eine Erläuterung zu den beiden unterschiedlichen Vorgehensweisen, bei der Erstellung des potenziellen Gebiets, in denen sich das Airbnb laut den Angaben aus Beschreibungen und Bewertungen befindet.

5.1 Erforderliche Vorarbeiten

Alle Informationen liegen im *World Geodetic System 1984* (WGS 84) vor. Dieses Referenzsystem verwendet als Standardeinheit das Grad. Die Berechnung eines räumlichen Buffers auf Basis einer Angabe in Metern ist nicht unmöglich, aber umständlich. Deshalb wird bei allen verwendeten Geodaten das Referenzsystem auf *österreichisches Bundesmeldenetz M34* geändert. Nun können Meter als Distanzangaben eingetragen werden.

Damit das in Kapitel 2.6 angewandte Verfahren zur Berechnung der Isochronen angewandt werden kann, sind räumliche Adaptionen zwischen den POIs, den Unterkünften und der GIP nötig. Die POIs und die Gaststätten sind auf einer Kante oder auf einen Knoten des Netzwerkes zu platzieren. Nur mit diesem Schritt ist es möglich, dass beim späteren Wege-Such-Verfahren, die POIs als Startpunkte und die Airbnb als Endpunkte fungieren können. Für dieses Vorhaben wurde das Werkzeug *Near* von ArcMap verwendet. Das Ziel dieses Werkzeug ist es, das Eingabefeature, innerhalb einer festgelegten Suchtoleranz, zum nächstgelegenen Nearfeature hinzuziehen. Es ist zu beachten, dass bevor das Werkzeug die Aufgabe abarbeitet, nur die relevanten GIP-Kanten zur Verfügung stehen. Ein dafür angewandter Filter soll alle Features selektieren, die keine Bundes- oder Landestraße sind, sowie nicht als Fußgängerzonen markiert sind. Somit wird ausgeschlossen, dass das U-Bahn-Netz, die Autobahnen oder die Eisenbahnschienen als Fußweg erkannt werden. Das ist sowohl in dieser Phase als auch später bei der Erstellung des Routing-Netzwerkes wichtig. Außerdem wurden für das Routingnetzwerk keine Höhenunterschiede berücksichtigt.

Die Distanzangaben aus den Bewertungen und Beschreibungen können sowohl in Metern als auch in Minuten vorliegen. Bei der räumlichen Angabe ist eindeutig, dass diese Entfernung eins-zu-eins weiter analysiert wird. Für die Analyse im GIS bedarf es bei einer zeitlichen Distanzangabe, einer Umrechnung des Wertes in Metern. Um die Annahme aus Kapitel 2.5 fortzuführen, dass sich ein Fußgänger mit einer Geschwindigkeit von fünf Kilometern pro Stunde bewegt, muss jeder zeitliche Wert mit 83,33 multipliziert werden, um einen Meter-Wert zu erhalten.

5.2 Auswahl der treffendsten Distanzangabe

Wie schon in Abbildung 18 erkennbar, ist es möglich, dass es zu einer Unterkunft mehrere Beschreibungen und Kommentare gibt, die auf das gleiche Objekt referenzieren. Im Besten Fall sind die Distanzangaben in den verschiedenen Fällen ident. Davon darf allerdings nicht ausgegangen werden. Es gilt also nun herauszufinden bzw. festzulegen, welche der Angaben in die Analyse mithineinfließt.

Bei einer Distanzangabe pro Referenzobjekt und Airbnb ist es logisch, dass eben nur diese eine herangezogen wird. Sind zwei unterschiedliche Distanzangaben vorhanden, muss eine Entscheidung getroffen werden, welche Information als gültig markiert wird und weiter in die GIS-Analyse übergeht. Die POIs weisen für Touristen eine hohe Attraktivität auf, immerhin haben die Personen, eben diese Orte in der Bewertung erwähnt. Sie waren ein wesentlicher Bestandteil während ihres Aufenthalts. In diesem Zusammenhang konnten Lloyd & Heivly 1987 feststellen, dass mit steigendem Abstand zwischen der Residenz und dem POI, die kognitive Distanz immer weiter überschätzt wird. Bei einer realen Distanz von 100m betrug die kognitive Distanz zwischen 100 und 170m und bei 600m bereits zwischen 620 und 700m. In der Studie wurden Personen befragt, die nicht in der Nähe des zentralen Stadtviertels wohnen. Diese Gruppe nutzt die Innenstadt auch vorwiegend touristisch. Nach dieser Überlegung unterschätzen Personen von Außerhalb die Distanzen. Deshalb wird in der GIS-Analyse, beim Auftreten zweier Distanzen, das Maximum genommen, um eine vermeintliche Fehleinschätzung des einen Touristen durch die Einschätzung des anderen auszugleichen. Es kann nicht davon ausgegangen werden, dass bei zwei kognitiven Distanzangaben immer das Maximum, die richtig Wahl ist. Doch in diesem Fall ist es die einzige, mit Literatur belegbare Annahme.

Bei drei und mehr Distanzinformationen pro Unterkunft und POI ergibt sich eine neue Möglichkeit, die vorher nicht anwendbar war. Um Ausreißer nach unten oder oben zu umgehen, bietet es sich an den Median – ganz im Gegenteil zum arithmetischen Mittel – aller Angaben zu wählen. (Wilfling 2011) Mit dieser Methode fließen alle angegeben Werte in die Analyse mit ein. Die Festlegung erfolgt dennoch auf einen Wert.

5.3 Ergebnisse

Die relevanten Datensätze wurden aus den unstrukturierten Daten extrahiert und alle vorhandenen Geodaten liegen im gleichen Referenzsystem vor. Die Schritte, welche vor der Analyse im GIS vollzogen werden müssen, sind erledigt. Es gilt zu überprüfen, ob die Auswertung mittels Luftlinien oder auf Basis eines Routing-Netzwerkes befriedigendere Ergebnisse liefert. Die Anwendung beider Verfahren soll einen Vergleich ermöglichen, um festzustellen, ob die Berücksichtigung eines Routing-Netzwerkes einen positiven Mehrwert für das Analyseergebnis hat. Ein solches Netzwerk stellt eine zusätzliche potenzielle Fehlerquelle (fehlende Stützpunkte, falsche Verbindungen oder Falscheinträge) dar.

5.3.1 Erstellung des Einzugsgebiets

Luftlinie

Bei der Berechnung der Luftlinie rund um ein Eingabefeature ist keine räumliche Verknüpfung mit anderen Features von Nöten. Als Eingabewerte dienen alle Distanzangaben, die nur noch in Meter vorliegen. Überschneiden sich zwei oder mehrere sogenannte Puffer, so stellt diese Fläche ein potenzielles Gebiet dar, in dem sich das jeweilige Airbnb befindet. Wenn es keine Schnittmenge gibt, so kann auch kein Potenzialgebiet ausgewiesen werden, da in einem solchen Fall kein Mehrwert durch das Verwenden von zwei oder mehr POIs entstanden ist. Daher ist für die nachfolgenden Messungen immer der nächste Überschneidungsbereich von mindestens zwei Puffern ausgewählt worden. Das gleiche gilt für die Auswertungen bei den Isochronen.

Netzwerkbasiert

Der wesentliche Unterschied in der Methodik zur vorherigen Analyse ist, dass die GIP nicht mehr irrelevant ist und für mehr als nur für Visualisierungszwecke herangezogen wird. In diesem Schritt wird ein Einzugsgebiet rund um die POIs gezogen, wobei die Impedanz, die berechnete Distanz in Metern ist. Es wird errechnet, wie weit eine bestimmte Meterangabe vom POI entfernt ist. Wenn sich zwei oder mehr Einzugsgebiete überschneiden, so ergibt sich daraus ein Potentialbereich. Das entspricht dem gleichen Prinzip von oben. Für die Erstellung des Netzwerkes und zur Messung von umgebenden Bereichen wurden die *Netzwerkfunktionen* von ArcMap angewandt. Es wird sich zeigen, ob die Implementierung der GIP einen solchen Mehrwert fürs Analyseergebnis beiträgt.

Der Unterschied der beiden Erstellungsmethoden ist in Abbildung 19 dargestellt. Das netzwerkbasierte Einzugsgebiet ist kleiner, da die Straßenverläufe als Barriere bei der zurückzulegenden Wegstrecke dienen. Dadurch wird das Ergebnis etwas realistischer, da die Schätzung der Nutzer auf eigenen Erfahrungen, etwa durch Selbstbegehung oder einer Navigationsapplikation basieren. Letztere verwenden unter anderem auch ein Routing-Netzwerk zur Berechnung von Distanzen und nicht Luftlinien. Im kleinen Ausschnitt der Abbildung 19 ist zu sehen, wie das Straßennetzwerk das Einzugsgebiet beeinflusst. Während es für die Luftlinie irrelevant ist, welche natürlichen oder künstlichen Hindernisse auftreten, wird das netzwerkbasierte Gebiet bei Nichtvorhandensein von Gehwegen beschnitten. Für diesen Airbnb-Standort zeigt sich des Weiteren, dass obwohl der Fußweg zum Schloss Schönbrunn zirka 30 Minuten lang ist, der Nutzer sich auf dieses Lokatum bezieht. Das wirkt zwar als lange Wegdistanz, ist aber nicht ganz unrealistisch, denn von den 42 relevanten zeitlichen Distanzangaben waren 31 gleich oder kleiner als 15 Minuten und 11 Angaben noch weiter entfernt. Das Maximum war eine Wegstrecke von 40 Minuten. Die Abbildung 19 bestätigt zumindest für diesen Fall, dass die Annahme für die Fußgänger-Geschwindigkeit von 5 km/h richtig gewählt ist, denn das relevante Airbnb befindet sich immer im Grenzbereich beider Einzugsgebiete.

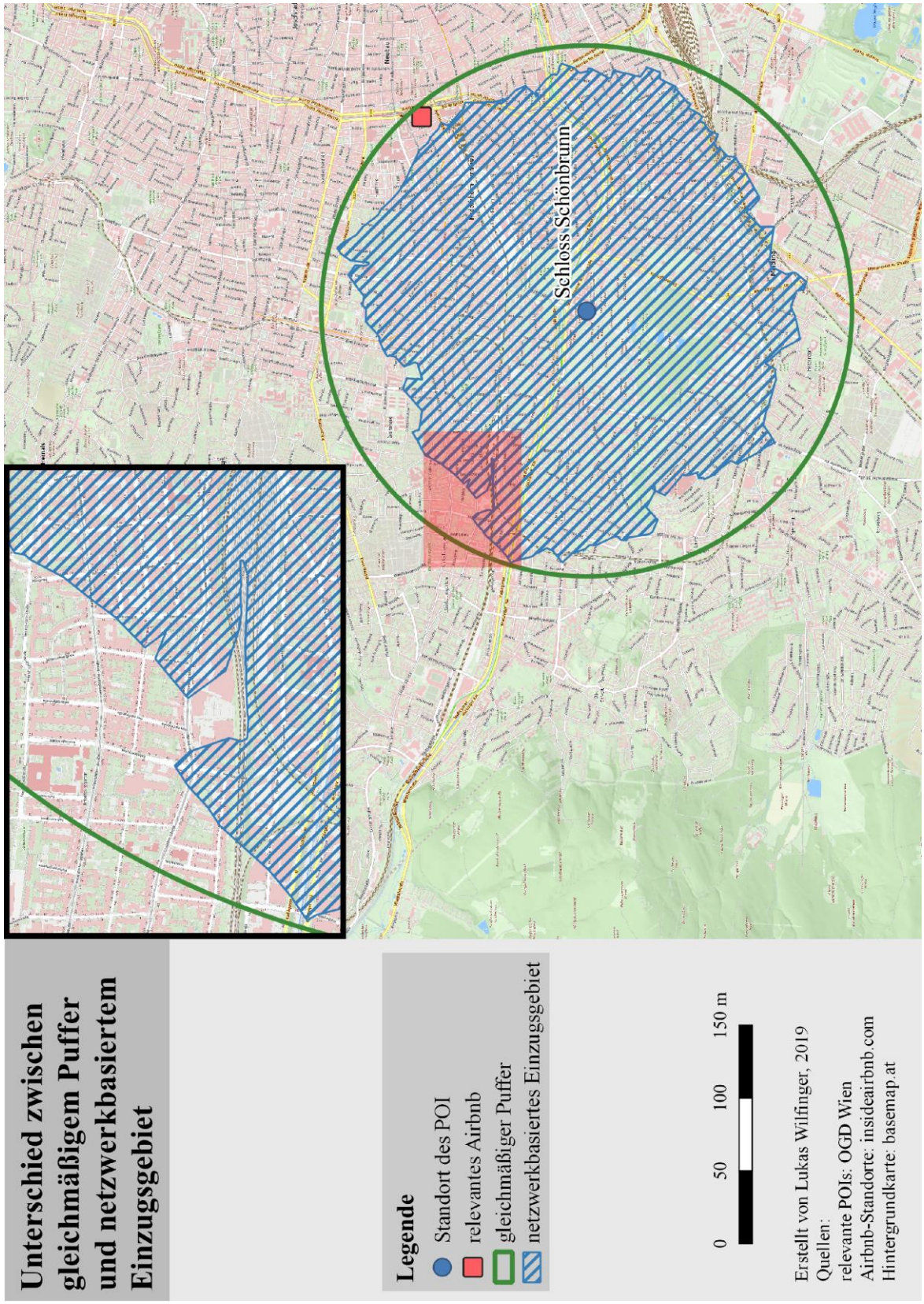


Abbildung 19 Unterschied zwischen Puffer und netzwerkbasiertem Einzugsgebiet

5.4 Messen und Vergleichen der Ergebnisse

Zur Klarstellung der Frage, ob es generell möglich ist, auf Basis von nutzergenerierten Informationen, einen Standort mittels relativer Angaben festzustellen, wird eine Distanzmessung eingesetzt. Dabei gilt es zum einen, die zwei unterschiedlichen Netzwerkanalysen zu betrachten (gleichmäßiger Puffer und netzwerkbasierte Isochrone), als auch zwei verschiedene Bezugspunkte für die Messung zu wählen. Diese Bezugspunkte sind die Distanz zum nächsten Grenzpunkt der Überschneidungsbereiches und der Entfernung zum Zentroid (geometrischer Mittelpunkt) dieser Fläche. Die Messung erfolgte händisch mit Hilfe des „Fangen"-Werkzeuges. Dabei erfolgt ein automatisches hinziehen zu einem Punkt oder Stützpunkt, wenn sich der Mauspfeil darüber befindet. Das Längenmaß der Distanzmessung ist Meter. Mit diesen beiden Messpunkten (nächster Grenzpunkt und Zentroid) soll auch festgestellt werden, dass obwohl der nächste Punkt des Potenzialbereichs beispielsweise nicht weit entfernt ist, es durch eine große Entfernung zum Zentroid einen hohen Unsicherheitsfaktor gibt. Ein solcher Fall würde darauf hindeuten, dass die potenzielle Fläche, in der sich das Airbnb befinden könnte, so groß ist, dass weitere Einschränkungen nötig sind. In Abbildung 20 ist Vorgehensweise bei der Messung dargestellt.

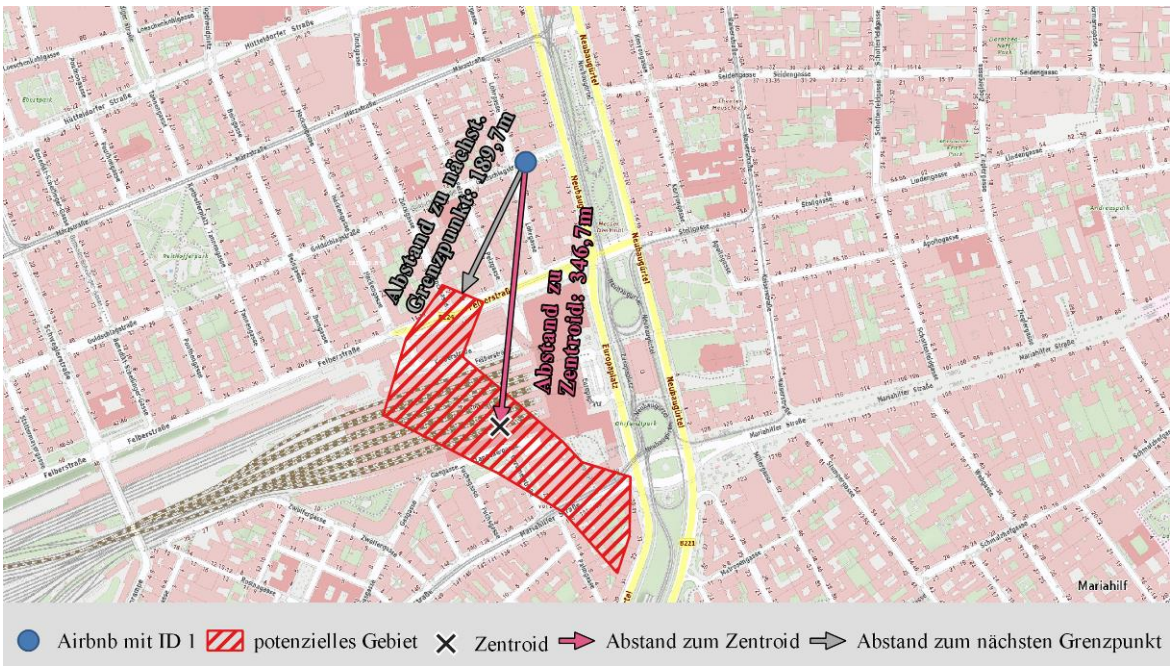


Abbildung 20 Beispielhafte Distanzmessung bei Airbnb ID 1 (Hintergrund: basemap.at)

In Tabelle 13 sind die Ergebnisse aus den Entfernungsmessungen aufgelistet. Wenn 0 Meter als Abstand eingetragen ist, so befindet sich das gesuchte Airbnb genau im Überschneidungsbereich. Um eben dennoch festzustellen, ob die Größe des entstandenen Suchkorridors stark eingeschränkt wurde, ist die Differenz zwischen dem nächsten Grenzpunkt und dem Zentroid angegeben. Außerdem ist die Anzahl an eindeutigen POIs angegeben, auf die sich die Berechnung des Überlappungsbereich stützt.

Werden nur die Entfernungen zu den nächsten Grenzpunkten betrachtet, so sind es bei den durch Puffer erzeugten Bereichen drei potenzielle Standorte, die genau in einen Überlappungsbereich fallen, während es bei den Isochronen nur zwei sind. Das allein ist aber noch kein Anspruch auf Präzession, denn es ist auch nicht zu unterschätzen, wie groß der Überschneidungsbereich ist. Ein Indiz für die Größe soll der Abstand vom realen Standort zum Zentroid darstellen. Je größer dieser Weg ist, desto stärker ist auch der Unsicherheitsfaktor bei der Suche nach der tatsächlichen Position. Für zusätzliche Erkenntnisse sorgt der Vergleich von Puffer- mit Netzwerkdistanzen. Ein besonders starker Unterschied ist beim möglichen Standort von Airbnb 1 zu sehen. Die Differenz zwischen den beiden Entfernungsangaben zum nächsten Grenzpunkt beträgt hier 154,3 Meter. Dieser Umstand ist vor allem auf das Vorhandensein von Bahngleisen zurückzuführen, auf die noch näher in Kapitel 5.5 eingegangen wird. Die größte Ungenauigkeit zeigt sich beim errechneten Suchbereich von Airbnb mit der ID 3. Der exakte Standort liegt zwar immer im Überschneidungsbereich, allerdings ist dieser im Vergleich zu den anderen Bereichen am größten. Am besten abgeschnitten hat jener Standort, in dem die meisten relevanten POIs genannt wurden. Bei Airbnb mit der ID 4 sind insgesamt sieben Distanzangaben zu unterschiedlichen Referenzpunkten angegeben worden. Daraus entstand eine maximale Ungenauigkeit von 50 m (Puffer) bzw. 122,6 m (Isochrone).

ID	**Puffer: Abstand in Meter zum**			**Isochrone: Abstand in Meter zum**			**Anz. POI**
	nächst. Grenzpunkt	Zentroid	**Differenz**	nächst. Grenzpunkt	Zentroid	**Differenz**	
1	35,4	287	**251,6**	189,7	346,7	**157**	2
2	0	257,4	**257,4**	21,7	243	**221,3**	5
3	0	300	**300**	0	318,7	**318,7**	4
4	0	50	**50**	0	122,6	**122,6**	7
5	108,2	208,2	**100**	106	190,3	**84,3**	5
7	13,8	212,6	**198,8**	49,6	209	**159,4**	2

Tabelle 13 Ergebnis der Messung der Entfernungen

Der Idealfall bei der Suche nach dem Potenzialbereich ist, wenn zu einem Airbnb maximal zwei unterschiedliche Distanzangaben zu finden sind. Das verhindert eine Berechnung des Medians und damit das subjektive Eingreifen in den Prozess. Es stellt sozusagen die „rohe“ Datenbasis dar. Dieser Fall ist bei dem Airbnb mit der ID 7 eingetreten. Die Bewertungen bezogen sich einmal auf den Stephansplatz und einmal auf die Landstraße Wien-Mitte. Die Unterkunft befindet sich, räumlich gesehen, zwischen den beiden POIs. Zwar liegt der reale Standort nicht innerhalb des Überschneidungsbereich, allerdings nahe an der Grenze. (Abbildung 21) Die restlichen Karten, auch jene bei denen die Einzugsgebiete mittels netzwerkbasierter Distanzen erstellt wurden, finden sich in Anhang IV.

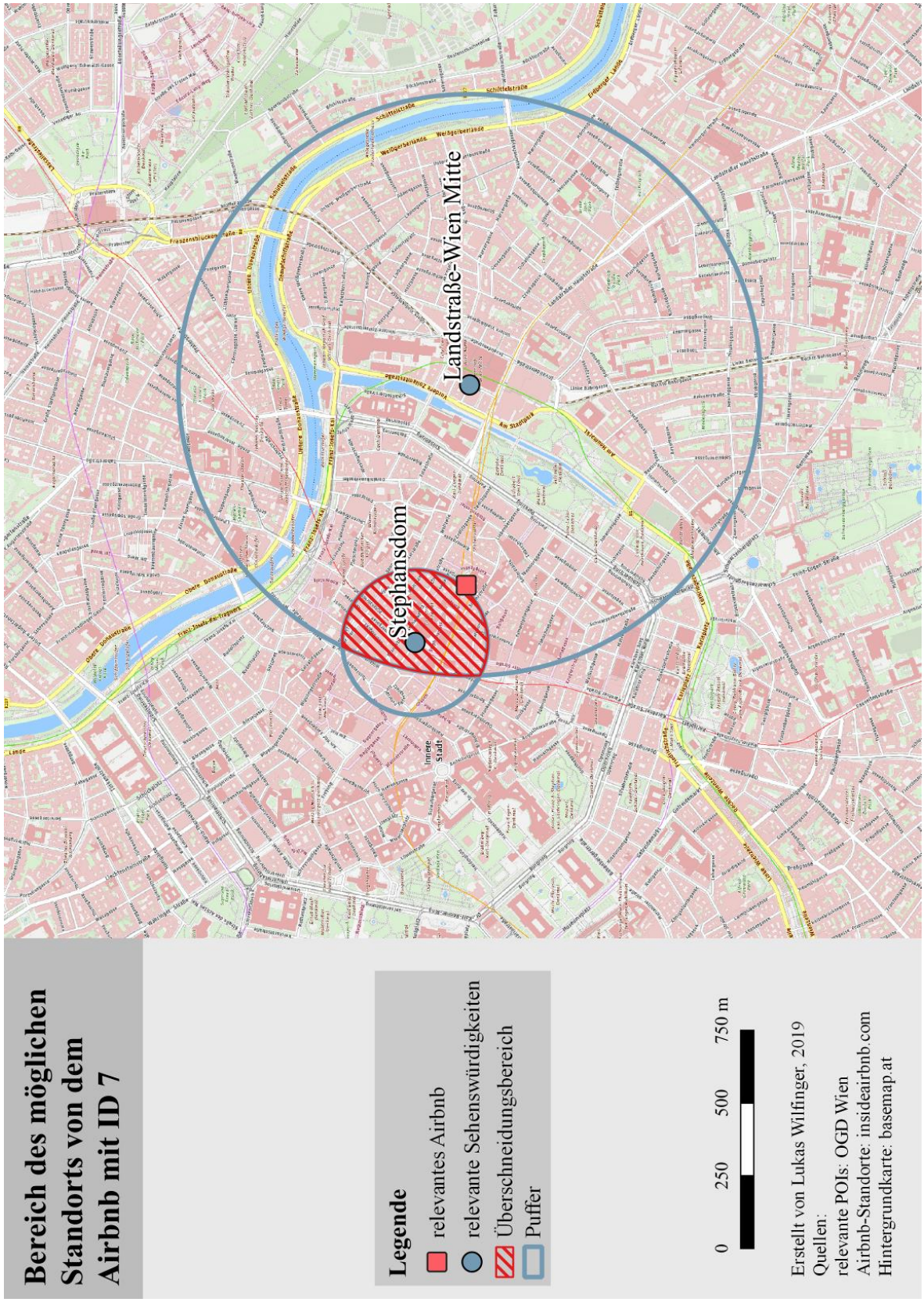


Abbildung 21 ideale Erstellung des Überschneidungsbereichs bei Airbnb mit ID 7

5.5 Bewertung

In sechs von insgesamt zwölf Fällen ist die Puffer-Distanz näher am möglichen Standort, als die Isochrone. Bei vier Vergleichswerten ist die Isochrone weniger weit entfernt und in zwei Fällen befindet sich der reale Standort der Unterkunft im Überschneidungsbereich. (Tabelle 13) Das lässt darauf schließen, dass die Puffer-Abstände genauer und zielführender sind. Es ist allerdings zu beachten, dass dieser Puffer ohne Rücksicht auf die realen Straßenverhältnisse erstellt wurden. Für die Analyse sind keine gesetzeswidrigen oder physisch unmöglichen Überquerungen berücksichtigt worden.

Bei der Bildung der Isochronen liegt hingegen, die GIP zugrunde. Diese verhindert, bei der korrekten Auswahl, dass es nicht möglich ist, durch Häuserblöcke hindurch zu gehen oder Bahngleise zu übertreten. Für die Auswertung als realistischer anzusehen ist die Distanz, die auch ein Routing-Netzwerk berücksichtigt hat. Damit das Ergebnis als noch valider einzustufen ist, wurde für die zwei Airbnbs mit der ID 1 und 4 der Überschneidungsbereich noch weiter verkleinert, indem ein Realnutzungsplan der Stadt Wien in die Analyse eingeflossen ist. Die Annahme dahinter ist, dass sich ein Airbnb nur in einem Wohnhaus befinden kann. Für eine solche Untersuchung bedeutet das, dass nur Flächen mit dem Attribut *Wohn- u. Mischnutzung (Schwerpunkt Wohnen)* mit der möglichen Standortfläche verschnitten wurden.

Eine solche beispielhafte Auswertung ist in der Abbildung 23 und der Abbildung 24 zu sehen. Bei ersterer ergab sich dadurch, dass das Suchgebiet weiter verkleinert werden konnte. In einem der schwarz karierten Polygone befindet sich demnach das Airbnb. Bei einem visuellen Vergleich mit Abbildung 22 zeigt sich, dass auch der richtige Häuserblock im eingeschränkten Gebiet ist. Eine Verortung des exakten Türeingangs ist nicht valide und kann auch mit den vorhandenen Daten von Airbnb nicht nachvollzogen werden. Dazu ist die räumliche Auflösung auf der Webseite zu ungenau. (Abbildung 22) In diesem Fall wird aber immerhin der Häuserblock lokalisiert, in dem sich das Airbnb befinden könnte. Für die Lokalisierung einer solchen Güte waren sieben unterschiedliche Distanzangaben zu POIs und eine subjektive Interpretation von Nöten. Das ist die größte Variation innerhalb aller sechs untersuchten POIs. Sowohl die Routing-, als auch die Luftliniendistanz konnten bei dieser Unterkunft das beste Ergebnis erzielen.

Abbildung 22 Screenshot: airbnb.com zu Unterkunft mit ID 4 (Hintergrund: google.com)

Die berücksichtigte Realnutzung kann einen visuellen Mehrwert zur Auswertung beitragen und führt dazu, dass der Suchbereich auf ein Minimum eingeschränkt wird. Dieser Umstand wird dann ersichtlich, wenn im potenziellen Bereich ein Industriebereich, ein Gewässer oder ähnlich unbewohnbares Gebiet liegt. Die Verwendung der Realnutzung kann helfen einen Wohnblock zu finden, in dem sich die Unterkunft befindet. In Abbildung 24 war es zwar nicht möglich, das stark abweichende Ergebnis von 346,7m auf 0m zu verringern. Allerdings hat sich die Entfernung zum nächsten Grenzpunkt von 189,7m auf 100,1m verkleinert. Dies bekräftigt die Annahme, dass der nächste detektierte Wohnblock innerhalb des Überschneidungsbereichs, die ideale Wahl ist und deshalb für die Messung herangezogen wird.

Alle Karten zu den Airbnbs befinden sich in Anhang IV. Diese Darstellungen beinhalten, für jede berücksichtige Unterkunft, die beiden Entfernungsarten. (ohne Realnutzung) Wenn mehr als zwei POIs zur Verfügung standen wird die Komplexität bei der Suche nach dem Potenzialgebiet sichtbar. Bei zumindest drei verschiedenen Distanzangaben ergibt sich nicht ein eindeutiger Überschneidungsbereich, sondern mehrere. Oftmals auch welche, die weit vom realen Standort der Unterkunft entfernt sind. Aufgrund dieser Umstände bedarf es entweder einer visuellen Interpretation oder einer tiefgreifenderen Analyse, welche Distanzangaben verworfen werden können. Ohne einer Orientierung an den realen Standort wäre eine eindeutige Auswahl des Überschneidungsbereichs, zumindest bei den Airbnbs mit den IDs 2 und 4 nicht möglich gewesen. Bei einer anderen Unterkunft empfiehlt sich der folgende Ansatz, und zwar, dass der Überschneidungspunkt der Entfernungsangaben als Potenzialstandort verwendet wird. Dies verhindert von vorherein die Unsicherheit durch ein größeres Überschneidungsgebiet. In einen anderen Fall ist wiederum zu beachten wie groß der Abstand zwischen der kleinsten und der nächstgrößten Entfernungsangabe ist. Deswegen kann die Hypothese aufgestellt werden, dass beim Überschreiten eines gewissen Schwellwerts, die minimale Angabe als korrekt angesehen werden kann. Die Erfahrungen aus der praktischen Auswertung sind in Tabelle 14 aufgelistet.

Erfahrungsgewinn aus praktischer Auswertung	**ID1**	**ID2**	**ID3**	**ID4**	**ID5**	**ID7**
Idealfall: Maximal zwei Distanzangaben	X					X
Überschneidungsbereich falscher Ansatz: Überschneidungspunkte besser geeignet			X			
Ohne visuelle Interpretation kein Ergebnis		X		X		
Großer Abstand zwischen kleinster und nächstgrößter Entfernungsangabe					X	

Tabelle 14 Lerneffekte nach der praktischen Auswertung

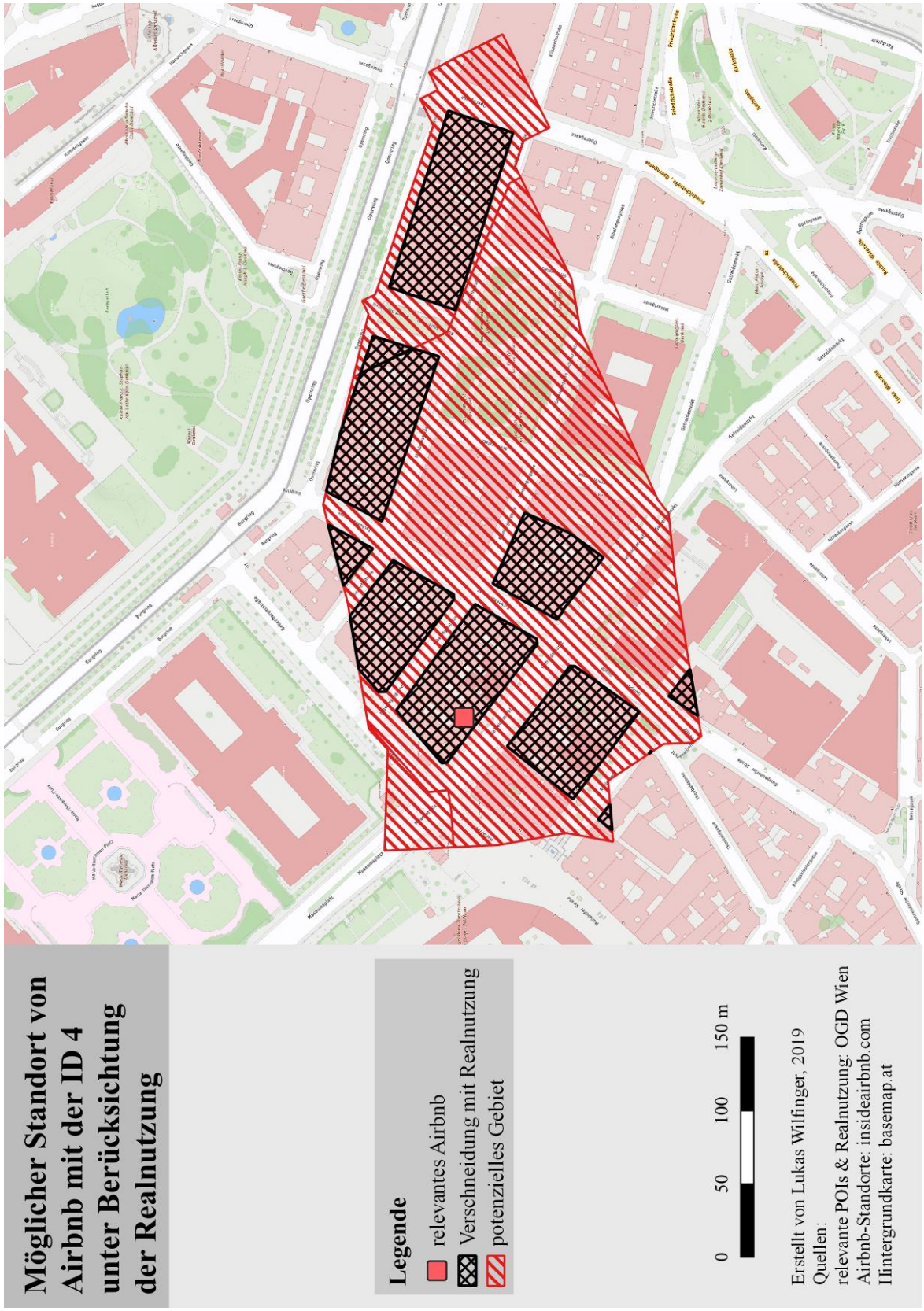


Abbildung 23 Berücksichtigung der Realnutzungskartierung bei Airbnb ID 4

Abbildung 24 Detektion von möglichen Wohnblöcken durch die Realnutzung

6 Diskussion

Am Anfang dieses Kapitel steht die Beantwortung der Forschungsfrage und deren Arbeitsfragen, die im Abschnitt 1.4 definiert sind. Begonnen wird mit jener Arbeitsfrage, welche durch eine umfangreiche Literaturrecherche beantwortet werden kann.

- *Wo befindet sich der Forschungsstand zum Thema ‚Geolokalisation auf Basis von Text Mining' und inwieweit können daraus folgende Erkenntnisse unterstützend auf diese Arbeit wirken?*

Die derzeitige Forschung im Umfeld von Geoparsing hat sich auf die Klärung von Unsicherheiten bei der NER mit Twitter oder anderen sozialen Medien als Datenquelle fokussiert. (Middleton et al. 2018; Utomo, Adji & Ardiyanto 2018; Karagoz et al. 2016) Es ist schwierig die erreichten Genauigkeiten von unterschiedlichen Studien zu vergleichen, da verschiedene Methoden, Sprachen und Granularitäten verwendet werden. Außerdem ist zu unterscheiden, was das Ziel der Forschung ist, denn zum einen kann die Position bestimmt werden, auf die sich eine Nachricht bezieht oder es ist wichtig von wo aus die Information abgesetzt wurde. (Hu 2018) Für diese Arbeit von Relevanz ist ersteres und das auf Basis einer NER mittels der Methode des *Gazetteering*. Die maximale Genauigkeit auf städtischer Granularität, die damit zurzeit erreicht wird, liegt bei 14%. Das ist nach Inkpen et al. 2017 keine Überraschung, denn der Gazetteer-basierte Ansatz ist zwar simpel, allerdings können alle unterschiedlichen Schreibweisen und Spitznamen für Orte nicht abgedeckt werden. Dieses schwache Ergebnis war, unter anderem, ein Anlass dafür, räumlich relevante Sätze über die metrischen Distanzangaben zu bestimmen. Das Gazetteering wurde darauffolgend nicht automatisiert vorgenommen, sondern erst nach erfolgter Normalisierung der POIs. Die Gründe dafür wurden in Kapitel 4.4.2 wiedergegeben. Es ließen sich keine vergleichbaren Studien finden, welche die metrischen Distanzen in den Texten in den Mittelpunkt gestellt haben. Das hat vor allem den Grund, da Menschen die Entfernungen meist in qualitativer Form verbal oder schriftlich wiedergeben. Diese Tatsache ist in Kapitel 2.4.2 dargelegt. Dieser Forschungsbereich widmet sich nicht nur der Quantifizierung von „nah" und „fern" (Chen, Vasardani & Winter 2016), sondern auch der Umlage von kognitiven auf räumlich-reale Gebiete. (Purves et al. 2018)

Die Beantwortung der obigen Arbeitsfrage hat geholfen die Hypothesen für die praktische Analyse festzulegen, da nun verständlich ist mit welchen Herausforderungen zu rechnen ist. Mittels der Definition von Anforderungen und Hypothesen sollen nur jene Datensätze in die Analyse miteinfließen, die einen räumlichen Kontext haben. Dies vermeidet mögliche unnötige Interaktionen in späteren Schritten.

- *Welche Hypothesen sind bei der Extraktion von räumlich relevanten Inhalten aus textuellen Informationen festzulegen, um eine sinnvolle Datenbasis für die Analyse im GIS zu erstellen?*

Die wichtigste Annahme für eine erfolgreiche räumliche Informationsextraktion aus Textdaten ist, dass die verwendete Datenquelle ein Potenzial dafür aufweist. Wenn das festgelegt ist, dann gibt es verschiedene technische und heuristische Ansätze die Daten GIS-tauglich zu machen. Ein möglicher Ansatz wurde in dieser Arbeit angewandt. Zunächst wurde angenommen, dass metrische Distanzen nur im Zahlenformat vorkommen. Auf die Entfernung folgt eine der in Tabelle 11 angeführten Einheiten oder deren Abkürzungen. Es wurde überprüft ob zwischen Zahl und Einheit ein Abstand ist oder nicht. Nur die Sätze, die diese Hypothese bestätigten, wurden in die händische Kodierung übergegeben. Dieser Schritt hatte zur Folge, dass der Korpus stark ausgedünnt wurde, und zwar von 9.107 auf 293 potenzielle Sätze. Dann wurden die Sätze je nach Aussagekraft kodiert. (s. Anhang III) Die Sätze mit der Kodierung „2“ enthalten Informationen, die in einem GIS zu verarbeiten sind. Bei dessen Vergabe musste auch die Annahme getroffen werden, dass Benutzer mit allen Synonymen von „Unterkunft“, „Airbnb“ und weiteren, jenen Ort meinen, an dem sie übernachtet haben und auf den sich die Bewertung oder Beschreibung bezieht. Aber auch die Personalpronomen „ich“ oder „wir“ können einen Aufschluss über dessen Position geben. Im folgenden Beispielsatz „wir befanden uns 300m von der Oper entfernt“ ist ersichtlich, dass sich das „wir“ auf den Standort des Airbnbs bezieht. Bei der händischen Kodierung wurden keine Annahmen gemacht, wenn es sich um andere Sprachen als Deutsch oder Englisch handelte. Dies geht auch aus den Anforderungen aus Kapitel 3.2.2 hervor. Zur Klärung der Genauigkeit der Benutzerangaben wurde auf zwei wesentliche Methoden eingeschränkt, die auch in Abschnitt 2.5 erläutert sind. Entweder geben Personen die Distanzen mehr als Schätzung auf Basis einer Mental Map ab oder sie fragen diverse Applikationen zur Navigation nach der exakten Distanz. Im Falle der netzwerkbasierten Distanzberechnung wurde angenommen, dass sich Personen an die geltende Straßenverkehrsordnung halten und keine Autobahnen oder Bahnschienen kreuzen. Auch wurde festgelegt, dass sich Fußgänger in einer Stadt relativ frei bewegen können.

Nachdem die Hypothesen für die praktische Auswertung abgesteckt und umgesetzt sind, soll nun die Forschungsfrage beantwortet werden. Es ist zu klären, ob sich die verwendete Datenquelle für eine solche Untersuchung eignet und welche Genauigkeit erreicht werden kann.

- *Ist es möglich den Standort auf Basis von relativen Entfernungsangaben innerhalb von nutzergenerierten textuellen Daten zu bestimmen?*

Diese Frage kann mit einem Ja beantwortet werden, allerdings muss definiert sein, welche Abweichungen noch akzeptabel sind. Wird davon ausgegangen, dass Airbnb bei einem Großteil der Unterkünfte die exakte Position in einem Radius von bis zu einem Kilometer versteckt, dann war der erstellte Abstand aufgrund der Bewertungen und Beschreibungen immer geringer. In drei (gleichmäßiger Puffer) bzw. zwei (netzwerkbasiertes Einzugsgebiet) Fällen befand sich das Airbnb innerhalb des gefundenen

Überschneidungsbereiches. Zur Erstellung des potenziellen Gebietes mit der niedrigsten Größe, flossen sieben Distanzangaben in die Berechnung mit ein. Es wurde demnach aus sieben Polygonen, der daraus entstehende Überschneidungsbereich selektiert. Mit dieser Erkenntnis kann das Potential, welches Middleton et al. 2018 festhielten, dass relativen Entfernungsangaben von nutzergenerierten Daten einen positiven Beitrag zur Standortverfeinerung haben, bestätigt werden. Mit der zusätzlichen Verwendung der Realnutzung konnte das Ergebnis soweit verfeinert werden, dass auch der richtige Häuserblock im Potenzialgebiet liegt, in dem sich das Airbnb befindet. (Abbildung 23) Mit dieser weiteren Datengrundlage ist es möglich den Überschneidungsbereich weiter einzugrenzen und irrelevante Bereiche, wie Bahnhöfe oder Parkgaragen auszuschließen. Den möglichen Standort auf wenige Meter einzugrenzen ist mit dieser Datengrundlage nicht möglich. Dazu sind die angegebenen POIs zu weitläufig und die daraus vermutete Position, welche der Verfasser wiedergibt, somit nur vage. Außerdem sind die Standorte der Airbnbs auf der Vermittlungswebseite selbst schwach aufgelöst, sodass maximal der richtige Häuserblock lokalisiert werden kann.

Während des Prozesses zur Findung von Standorten auf Basis von relativen Entfernungsangaben haben sich Problemstellungen ergeben, die es mit der letzten Arbeitsfrage aufzulisten gilt.

- *Welche praktischen Problemstellungen ergeben sich bei der Aufbereitung und Verarbeitung von textuellen Daten mit Raumbezug für ein Geoinformationssystem?*

Von den 9.107 analysierten Sätze beinhalteten nur 0,65% einen räumlichen Ausdruck mit quantitativer Information. Den sozialen Netzwerken werden doch Anteile von 2% zugeordnet, allerdings ohne genaue Distanzangaben. Je mehr potenzielle Distanzangaben im Text zum Analysieren vorhanden wären, desto exakter könnte die Aussage zu der treffendsten Beschreibung fallen. In den meisten Fällen wird nur eine Distanz zu einem POI angegeben. Das schafft eine hohe Abhängigkeit, dass eben diese eine Beschreibung stimmt. Wenn das nicht eintrifft, dann ergeben sich eventuell keine Überschneidungsbereiche und der gesuchte Standort kann nicht abgeleitet werden. Das gleiche Problem wird auch ausgelöst, wenn die angenommene Gehgeschwindigkeit nicht der Realität entspricht. Der Idealfall wäre, wenn alle Bewertungen einer Person auf der Plattform analysiert werden, um dessen Geschwindigkeit ableiten zu können. Eventuell überprüft die Person dessen Angaben auch immer mit dem identen Navigationsdienst. Die geringe Datenmenge ist auch generell ein Problem, da von den 13 ausgewählten Airbnbs nur sechs zur Beantwortung der Fragestellung beitragen konnten. Es hat sich demnach auch gezeigt, dass die Anzahl an Kommentaren und Bewertungen keinen Einfluss auf die Anzahl an räumlichen Ausdrücken mit quantitativen Angaben hat. (Abbildung 17) Die geringe Datenmenge ist möglicherweise auch der restriktiven Filterung von potenziell räumlich relevanter Information geschuldet. Wenn eine Zahl ausgeschrieben war, dann ist diese nicht detektiert worden. Außerdem haben es Sätze, die in einer anderen Sprache als Deutsch oder Englisch verfasst waren, nicht in die Analyse geschafft. Ein weiteres praktisches Problem bei der Kodierung stellte die unterschiedliche Schreibweise von POIs dar. In Kapitel 4.4.2 sind elf verschiedene Ausprägungen des „Wiener Hauptbahnhofs“ zu lesen. Durch diesen Umstand musste doch händisch in den Prozess zur Extraktion von

räumlich relevanter Information eingegriffen werden. Dazu findet sich in Anhang III die bearbeitete Tabelle. Die Problemstellungen beschränken sich zusammenfassend auf die orthographischen Fehler, die bei nutzergenerierten Daten nicht neu sind (Tabelle 3) und auf eine geringe Datenmenge, die für die Anwendung relevant waren.

Im Forschungsbereich des Geoparsings wird immer wieder Twitter als Datenquelle verwendet. Dies erlaubt es den Forscherinnen und Forschern, hunderttausende an Sätze zu untersuchen und auf deren räumliche Relevanz zu überprüfen. Außerdem beschränkt sich die Suche meist auf Toponyme. Deren Kontext wird durch ein zeitlich determiniertes Ereignis festgestellt. Beispiele hierfür sind Naturkatastrophen oder Terroranschläge. (Karimzadeh et al. 2019; Middleton & Krivcovs 2015) Der Kontext ist aber eben nicht das Wesentliche, ganz im Gegenteil zu dieser Arbeit. In diesem Fall sind sowohl die Toponyme als auch deren ‚umgebende' Wörter von Bedeutung. Die händisch kodierten Sätze sind im Anhang III aufgelistet und, vor allem die Sätze mit der Kodierung 1 sollen zeigen wie schwierig es für einen Computer sein könnte, die richtige Information zu extrahieren. Durch die kleinere Menge an Daten, verglichen mit Tweets, ist ein erforderlicher großer Trainingsdatensatz für statistische Modelle herausfordernder. Nichtsdestotrotz ist die wissenschaftliche Belegbarkeit durch die Ausarbeitung in Kapitel 4.2.1 gegeben.

Ein wesentliches Problem in dieser Arbeit war es, die verschiedenen Distanzangaben bei Mehrfachnennung eines POI zum gleichen Airbnb richtig zu einer Angabe zusammenzufassen. In Kapitel 5.2 ist erläutert, dass je nach Anzahl an relativen Entfernungsangaben, eine andere Methode zur Zusammenführung herangezogen wird. Dessen ungeachtet bleibt ein Unsicherheitsfaktor vorhanden, ob nicht doch eine völlig richtige Angabe mit dieser restriktiven Methode weggefallen ist. In zukünftigen Arbeiten wäre zu berücksichtigen, inwieweit sich die Distanzangaben von Gästen und Gastgebern unterscheiden und welche Informationen am ehesten der Wahrheit entsprechen. Außerdem gilt es abzuklären, ob eine allumfassende Analyse von Geolokalisation und Sentiments innerhalb von Kommentaren eine sinnvolle Ergänzung zur Wahl von Distanzangaben darstellt. Eine umfassende Sentimentanalyse (auf Basis von bestimmten Wörtern in einem Satz, wird versucht die Gefühle des Autors, während dem Schreiben, zu deuten) von Airbnb-Daten wurde bereits von Cheng & Jin 2019 durchgeführt, während Ke 2017 auf eine generell feinere räumliche Untersuchung von Standorten verweist. Eine Verschneidung dieser zwei verschiedenen Forschungsmethoden würde zum Beispiel Hypothesen ergeben, die testen sollen, ob ein Kommentar mit einem negativen Sentiment-Score eine eher falsche Distanzangabe enthält.

Die praktisch angewandte Methodik folgte einem festgelegten Ablauf. Beginnend beim Download der Daten für Wien von Inside Airbnb, über die Verarbeitung in R, bis hin zur räumlichen Analyse in ArcMap und QGIS. Besonders der zweite Schritt erfordert, bei einer Umstellung auf eine völlige Automatisierung, einen erhöhten zeitlichen Aufwand. Der erste Schritt, ein Prozess zur Suche nach den räumlich relevanten Inhalten, ist bereits erfolgt. Die in Anhang I und Anhang II enthaltenen R-Skripte erlauben es, durch Hinzufügen oder Entfernen von IDs, die zu analysierenden Daten auf eine nachvollziehbare Weise zu erweitern. Die im Anschluss durchgeführte Kodierung der

einzelnen Sätze, um deren Wertigkeit für die weiteren Auswertungen zu bestimmen, wurde noch händisch durchgeführt. Dabei handelt es sich um den einfachsten, aber auch den zeitintensivsten Ansatz. Daraus ergab sich dann auch der sogenannte „*Gold Standard*". Dieser stellt einen hochqualitativen annotierten Korpus dar, der für weitere Untersuchungen als Testdatensatz herangezogen werden kann. (Wissler et al. 2014) Mit dieser Basis ist eine Automatisierung dieses Schrittes zu empfehlen, um eine quantitativ höhere Stückzahl an Airbnbs zu erreichen. Bei der Erstellung eines solchen Algorithmus, sei auf die Problemstellungen aus Kapitel 4.4.2 verwiesen.

Im Fall von 51 Sätzen mit einer unklaren Georeferenzierung (Code 1), hat sich der Autor auf das Zentrum der Stadt bezogen. Dabei ist unsicher was als Stadtzentrum erkannt wird. Die Wiener Innenstadt ist durch die „Ringstraße" von den anderen Wiener Gemeindebezirken getrennt. Entlang dieser Abgrenzung befinden sich viele Sehenswürdigkeiten, wie die Oper, der Heldenplatz oder das Burgtheater. Doch das Zentrum der Innenstadt bildet der Wiener Stephansdom. Nun ist im folgenden Beispielsatz unklar, wo sich das Zentrum befindet bzw. was damit gemeint ist. „*About 15 minutes walk to the centre.*" Es gibt zwei Möglichkeiten, die als realistisch anzusehen sind. Entweder die Touristen empfinden das Übertreten der Wiener Ringstraße als Eintritt in das Zentrum oder erst das Stehen auf dem Stephansplatz. Diese Unsicherheit hat dazu geführt, dass doch 17,4% aller Sätze als nicht relevant markiert werden konnten, obwohl jeweils eine Distanzangabe vorhanden war. Die restlichen Beschreibungen und Bewertungen mit diesem Code enthalten Weglängen zu räumlich undefinierten Cafés, Lebensmittelmärkten, Parkgaragen oder Haltestellen des öffentlichen Verkehrs.

7 Zusammenfassung und Ausblick

Diese Arbeit beschreibt zu Beginn, dass es einen wissenschaftlichen Bedarf dafür gibt, die räumliche Information aus Tripletten mit quantitativer Entfernungsangabe zu extrahieren. Das Analyseergebnis stellt nicht nur eine weitere Datenquelle für Geoinformationssysteme im Allgemeinen dar, sondern bietet auch Hilfsorganisationen in Notsituationen einen Mehrwert. Die Angabe von Entfernungen zu umliegenden, eventuell markanten Gebäuden oder Geländepunkten kann einen positiven Beitrag zur relativen Verortung leisten. Die derzeitige Forschung beschränkt sich meist auf die Suche von Toponymen mittels Techniken aus der NER oder widmet sich der semantischen Deutung von Tripletten mit qualitativem Informationsgehalt. Mit Airbnb wurde eine Datenquelle verwendet, die gegenüber Twitter oder Flickr ein höheres Potenzial zur Findung von metrischer Information beinhaltet. Die Lage der Wohnung und die damit verbundenen Distanzen von der Unterkunft weg, hin zu den POIs haben eine wesentliche Bedeutung für die Buchung.

In Abschnitt 2 wurde der derzeitige Stand der Forschung in mehreren relevanten Themenfeldern mittels Literaturrecherche aufgearbeitet. Das war zum einen das Data Mining, als übergeordneter Begriff, der die gesamte Arbeit im methodischen Sinn begleitete. Das Ziel von Data Mining ist immer eine Informations-/Wissensextraktion – wie auch im praktischen Teil dieser Thesis. Es folgte die Erläuterung der NLP und dessen wissenschaftliche Begründung. Den Weg ebnete damals Chomsky 1956 mit seinen Überlegungen zur Universalgrammatik, die ein Computer erlernen kann. Bei der NLP handelt es sich um einen Bestandteil des Textminings, welches methodologisch den Übergang von Text zu Raum bildet. Nach diesem Abschnitt wurde festgelegt, dass nur Sätze mit einer expliziten Verortung nach Hu 2018 von Relevanz sind. Mit diesem Wissen wurden die Fachtermini zur praktischen Extraktion von räumlich relevanter Information erläutert. Darin enthalten ist auch der in Abbildung 7 visualisierte Prozess, der dies ermöglicht und allgemeine Gültigkeit hat. Für die spätere praktische Arbeit sind vor allem die Herausforderungen, die andere Autoren schon hatten, von Interesse. Das sind zum einen orthographische Unstimmigkeiten oder Kodierungsfehler, aber auch mangelhafte Gazetteer und Unsicherheiten bei der Bestimmung von Entitäten. Den Abschluss bildeten die kognitiven Aspekte bei der Schätzung von Distanzen und die theoretische Grundlage zur Berechnung von Erreichbarkeiten.

Die verwendeten Datenquellen wurden in Kapitel 3 aufgearbeitet. Dabei ist wichtig anzumerken, dass der Urheber der Daten das Unternehmen Airbnb ist. Die Aufbereitung der Daten erfolgte jedoch über die Onlineplattform Inside Airbnb. Es ist nicht explizit festgeschrieben, dass es verboten ist, die Daten zu analysieren, aber auch nicht erlaubt. Ein rechtliches Problem stellt vor allem das *Scraping* der Daten dar. Es handelt sich dabei um einen automatisierten Prozess, bei dem Information von einer Internetseite heruntergeladen wird. Inside Airbnb wird auch in anderen wissenschaftlichen Arbeiten als Quelle genannt – vor allem immer dann, wenn nicht auf die Anzahl an Übernachtungen oder auf Buchungspreise eingegangen wird. Diese Informationen von Inside Airbnb sind lt. Airbnb auch „*schlichtweg falsch*“. Es handelt sich bei denen in dieser Arbeit ausgewerteten Daten

um nicht-sensible Firmendaten und somit wird angenommen, dass dessen Analyse rechtlich erlaubt ist. Der Rest des Kapitels bezieht sich auf die Aufbereitung der Daten und der Festlegung von Anforderungen, welche die Beschreibungen und Bewertungen erfüllen müssen, um in die Analyse im GIS miteinzufließen. Außerdem wurden die anderen Datenquellen vorgestellt, die frei verfügbar sind und auch für kommerzielle Zwecke eingesetzt werden können.

Abschnitt 4 erläuterte die eigens erstellte Vorgehensweise für die praktische Aufbereitung. Darin enthalten ist auch die Absteckung des Untersuchungsraumes – 13 Airbnbs aus Wien wurden für die Analyse herangezogen. (Tabelle 8) Des Weiteren wurden die verwendeten Technologien bzw. Softwares erklärt und mit deren jeweiligen Marktbegleitern gegenübergestellt. Für die praktische Aufbereitung erfolgte eine Erläuterung der wichtigsten Paradigmen und Funktionen aus R, sowie die Klarstellung der Hypothesen zur Extraktion der relevanten Tripletten. Eine dieser Annahmen war zum Beispiel, dass die metrischen Distanzen immer als Ziffern- bzw. Zahlensymbole und nicht als ausgeschriebene Zahl angegeben sind. Ohne solche Hypothesen wäre es nicht möglich diesen heuristischen Ansatz in die Praxis umzusetzen, da sonst die Anhaltspunkte gefehlt hätten, wie in bestimmten Situationen vorzugehen ist. Durch die vorgenommene Extraktion konnten nicht alle Airbnbs zur Analyse ins GIS überführt werden, da keine oder nur eine Entfernungsangabe zu einem POI vorhanden war. (Abbildung 17)

In Kapitel 5 finden sich alle notwendigen Vorarbeiten für die räumliche Analyse. Dazu zählt die Festlegung des Referenzsystems auf das *österreichisches Bundesmeldenetz M34*. Außerdem wurden alle Angaben, die als Minuten vorlagen in Metern umgerechnet. Dies erfolgte durch eine Multiplikation mit dem Faktor 83,33 (es entspricht damit einer Geschwindigkeit von 5 km/h). Des Weiteren wurde bei Mehrfachnennung eines POIs beim identen Airbnb festgelegt, dass der Median als Maß für die Distanz herangezogen wird. Die Auswertung der Ergebnisse zeigte, dass es bei zwei Entfernungsangaben durchaus möglich ist, den Potenzialkorridor stark zu minimieren. Die Berücksichtigung der generalisierten Realnutzung der Stadt Wien konnte den Bereich weiter einschränken, da Straßen, Museen und ähnliches unbewohnbares Gebiet ausgeschlossen werden kann. So konnte erreicht werden, dass sich der für diese Unterkunft relevante Häuserblock im Potenzialgebiet befand. (Abbildung 23) Jedoch ist zu sagen, dass es bei mehr als zwei Angaben zu vielen Überschneidungsbereichen kommen kann, sodass die in Tabelle 14 angeführten Erfahrungen bzw. Empfehlungen zu beachten sind.

Im 6. Abschnitt wurden abschließend die definierte Forschungsfrage und deren Arbeitsfragen beantwortet, als auch Aufschluss darüber gegeben, wie die praktische Umsetzung genauer und automatisierter vonstattengehen kann.

Die derzeitige Forschung in der NER oder in der NLP benutzt meist dieselben Datenquellen, wie Twitter oder Wikipedia. Dadurch ist das Trainingsset, das zum Trainieren von statistischen Modellen eingesetzt wird, meist ident. Durch diesen Umstand treten immer wieder die gleichen Probleme auf. In Zukunft sollten *reproduzierbare Anwendungen, Algorithmen und Datensets öffentlich zur Verfügung* gestellt werden, um Vergleiche zu ermöglichen und damit die Forschung zur Gewinnung von Geoinformation

aus Texten voranzutreiben. (Purves et al. 2018: 286) Des Weiteren gilt es den räumlichen Fußabdruck in einem Satz eindeutiger zu bestimmen. Es ist demnach zu klären, ob eine Linie oder eine Fläche ein Toponym besser repräsentiert, als ein Punkt. Außerdem haben die Geo-text Daten nur dann ihre Sinnhaftigkeit, wenn eine Verbindung von Text und Raum besteht. Es sind neue Methoden zur computergestützten Erkennung von räumlichen Beziehungen zu entwickeln. Die derzeitigen Geosemantiken decken nicht ab, ob eine Person *über* einen Ort oder *von* einem Ort aus schreibt. (Hu 2018)

8 Literatur

Aggarwal, Charu. 2015. Data Mining - The Textbook. Springer Cham Heidelberg New York Dordrecht London.

Airbnb. 2019a. Über uns. https://press.airbnb.com/de/about-us/ (7 January, 2019).

Airbnb. 2019b. Fast Facts. https://press.airbnb.com/fast-facts/ (10 February, 2019).

Airbnb. 2019c. Terms. https://www.airbnb.at/terms#eusec14 (10 June, 2019).

Airbnb. 2019d. Robots.txt. https://www.airbnb.at/robots.txt (10 June, 2019).

Airbnb. 2019e. AGB. https://www.airbnb.de/terms#eusec10 (1 February, 2019).

Attard, Judie. 2016. Strategies and Approaches for Exploiting the Value of Open Data. Rheinischen Friedrich-Wilhelms-Universität Bonn.

Baldwin, Timothy, Paul Cook, Marco Lui, Andrew MacKinlay & Li Wang. 2013. How Noisy Social Media Text, How Diffrnt Social Media Sources? *Sixth International Joint Conference on Natural Language Processing*, 356–364. Nagoya, Japan: Asian Federation of Natural Language Processing.

Barnes, Grenville. 1995. Land and Geographic Information Systems. In Russell C. Brinker & Roy Minnick (eds.), *The Surveying Handbook*. Springer US. doi:10.1007/978-1-4615-2067-2.

Becker, Richard A. A brief history of S. *AT&T Bell Laboratories*. doi:10.1080/09627250608553013.

Bey, Sören. 2014. Alignment of structured and unstructured data for decision support. *IBA Bachelor Thesis Conference*.

Bhadani, Abhay & Dhanya Jothimani. 2017. Big Data: Challenges, Opportunities and Realities. *CoRR* 1–24. http://arxiv.org/abs/1705.04928.

Blum, Norbert. 2004. *Algorithmen und Datenstrukturen*. Oldenbourg Verlag.

Bowen, Glenn A. 2009. Document Analysis as a Qualitative Research Method. *Qualitative Research Journal* 9(2). 27–40. doi:10.3316/qrj0902027.

Briggs, Ronald. 1973. Urban cognitive distance. *Image & environment: Cognitive mapping and spatial behavior.*, 361–388. New Brunswick, NJ, US: AldineTransaction.

Brinkhoff, Thomas & Hans-Peter Kriegel. 1994. The Impact of Global Clustering on Spatial Database Systems. *Proceedings of the Twentieth International Conference on Very Large Data Bases*.

Brittain, Jim, Mariana Cendon, Jennifer Nizzi & John Pleis. 2018. Data Scientist's Analysis Toolbox: Comparison of Python, R, and SAS Performance. *SMU Data Science Review* 1(2).

Brockhaus. 2018. Qualität (Wirtschaft). NE GmbH | Brockhaus.

Brüggemann, Holger & Peik Bremer. 2015. *Grundlagen Qualitätsmanagement: Von den Werkzeugen über Methoden zum TQM*. doi:10.1007/978-3-658-09221-4.

Chen, Hao, Maria Vasardani & Stephan Winter. 2016. Geo-referencing Place from Everyday Natural Language Descriptions. *ACM Trans. Spatial Algorithms Syst.* V. 1–29. doi:http://dx.doi.org/10.1145/0000000.0000000.

Chen, Hao, Stephan Winter & Maria Vasardani. 2018. Georeferencing places from collective human descriptions using place graphs. *Journal of Spatial Information Science* 17. 31–62. doi:10.5311/josis.2018.17.417.

Cheng, Mingming & Xin Jin. 2019. What do Airbnb users care about? An analysis of online review comments. *International Journal of Hospitality Management*. Elsevier 76(April 2018). 58–70. doi:10.1016/j.ijhm.2018.04.004.

Chomsky, Noam. 1956. Three models for the description of language. *IRE Transactions of Information Theory* 2(3). 113–124. doi:10.1109/TIT.1956.1056813.

Cios, Krzysztof, Witold Pedrycz, Roman Swiniarski & Lukasz Kurgan. 2007. *Data Mining - A Knowledge Discovery Approach*. Springer US.

Cleve, Jürgen & Uwe Lämmel. 2016. *Data Mining*. 2. Auflage. Berlin: Boston: De Gruyter Oldenbourg.

Cooney, Michael. 2018. Cisco predicts nearly 5 zettabytes of IP traffic per year by 2022. *Network World*. https://www.networkworld.com/article/3323063/cisco-predicts-nearly-5-zettabytes-of-ip-traffic-per-year-by-2022.html (13 June, 2019).

Cox, Murray. 2018a. Inside Airbnb - Adding Data to the debate. *Get the data*. http://insideairbnb.com/get-the-data.html (13 October, 2018).

Cox, Murray. 2018b. Inside Airbnb - Adding Data to the debate. *Behind*. http://insideairbnb.com/behind.html (10 June, 2019).

CRAN. 2019. CRAN package repository. https://cran.r-project.org/web/packages/ (22 June, 2019).

Cullinan, J, S Hynes & C O'Donoghue. 2008. Estimating Catchment Area Population Indicators Using Network Analysis. *Irish Geography* 41(3). 279–294.

Curtin, Kevin M. 2007. Network Analysis in Geographic Information Science: Review, Assessment, and Projections. *Cartography and Geographic Information Science* 34(2). 103–111. doi:10.1559/152304007781002163.

Dale, Robert. 2010. Classical approaches to natural language processing. In Nitin Indurkhya & Fred J. Damerau (eds.), *Handbook of natural language processing*, 3–7. second edi. Boca Raton, Florida: CRC Press, Taylor & Francis Group.

Data.gv.at. 2019. Zielsetzung. https://www.data.gv.at/infos/zielsetzung-data-gv-at/ (12 February, 2019).

Delozier, Grant, Jason Baldridge & Loretta London. 2015. Gazetteer-Independent Toponym Resolution Using Geographic Word Profiles. *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence Gazetteer-Independent*, 2382–2388. doi:10.1016/j.celrep.2016.09.063.

DeMers, M N. 2008. *Fundamentals of Geographic Information Systems*. Fourth edi. Chichester: Wiley.

Earle, Paul S., Daniel C. Bowden & Michelle Guy. 2011. Twitter earthquake detection: Earthquake monitoring in a social world. *Annals of Geophysics* 54(6). 708–715. doi:10.4401/ag-5364.

Eftimov, Tome, Barbara Koroušić Seljak & Peter Korošec. 2017. A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations. *PLoS ONE* 12(6). doi:10.1371/journal.pone.0179488.

Egenhofer, Max J. & Robert D. Franzosa. 1991. Point-set topological spatial relations. *International Journal of Geographical Information Systems* 5(2). 161–174. doi:10.1080/02693799108927841.

Egenhofer, Max J, David M Mark & John Herring. 1994. *The 9-Intersection: Formalism and its use for natural-language spatial predicates*.

Fazal, Shahab. 2008. *GIS Basics*. New Delhi: New Age International.

Fegeas, Robin G., Janette L. Cascio & Robert A. Lazar. 1992. An overview of FIPS 173, the spatial data transfer standard. *Cartography and Geographic Information Systems* 19(5). 278–293. doi:10.1080/152304092783762209.

Forbes. 2018. As A Rare Profitable Unicorn, Airbnb Appears To Be Worth At Least $38 Billion. *Forbes*. https://www.forbes.com/sites/greatspeculations/2018/05/11/as-a-rare-profitable-unicorn-airbnb-appears-to-be-worth-at-least-38-billion/#4d1419482741 (10 February, 2019).

Frey, Harald. 2015. Grundlagen der Fußverkehrsplanung. TU Wien - Institut für Verkehrswissenschaften.

Friedrich, C. 2014. Comparison of ArcGIS and QGIS for applications in sustainable spatial planning. Universität Wien.

Gahegan, Mark. 1995. Proximity operators for qualitative spatial reasoning. In Andrew U Frank & Werner Kuhn (eds.), *Spatial Information Theory A Theoretical Basis for GIS*, 31–44. Berlin, Heidelberg: Springer Berlin Heidelberg.

Gelernter, Judith & Shilpa Balaji. 2013. An algorithm for local geoparsing of microtext. *GeoInformatica* 17(4). 635–667. doi:10.1007/s10707-012-0173-8.

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

GIP. 2018. Rückblick Symposium - 10 Jahre GIP in Österreich. http://www.gip.gv.at/home.html (12 February, 2019).

Goddard, Cliff & Andrea C. Schalley. 2010. Semantic Analysis. In Nitin Indurkhya & Fred J. Damerau (eds.), *Handbook of Natural Language Processing:*, 94–114. second edi. Boca Raton, Florida: CRC Press, Taylor & Francis Group.

Golledge, Reginald G. & Robert J. Stimson. 1998. Spatial Behavior: A Geographic Perspective. *Economic Geography* 74(1). 83–85. doi:10.1126/science.11.277.620.

Goodchild, Michael F., May Yuan & Thomas J. Cova. 2007. Towards a general theory of geographic representation in GIS. *International Journal of Geographical Information Science* 21(3). 239–260. doi:10.1080/13658810600965271.

Gorunescu, Florin. 2011. *Data Mining - Concepts, Models and Techniques*. Springer-Verlag Berlin Heidelberg.

Grishman, Ralph & Beth Sundheim. 1996. Message Understanding Conference - 6: A Brief History Ralph. *COLING '96 Proceedings of the 16th conference on Computational linguistics - Volume 1*. doi:10.3115/992628.992709.

Gritta, Milan, Mohammad Taher Pilehvar, Nut Limsopatham & Nigel Collier. 2018. What's missing in geographical parsing? *Language Resources and Evaluation*. Springer Netherlands 52(2). 603–623. doi:10.1007/s10579-017-9385-8.

Guttentag, Daniel. 2016. Why tourists choose Airbnb: a motivation-based segmentation study. *Dissertation*. University of Waterloo.

Hammarberg, K., M. Kirkman & S. de Lacey. 2016. Qualitative research methods: when to use them and how to judge them. *Human Reproduction* 31(3). 498–501. doi:10.1093/humrep/dev334.

Hass, Kara. 2009. Measuring accessibility of regional parks: a comparison of three GIS techniques. San Jose State University.

Hecht, Brent & Darren Gergle. 2010. On the " Localness " of User-Generated Content. *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, 229–232.

Henten, Anders Hansen & Iwona Maria Windekilde. 2016. Transaction costs and the sharing economy. *Info: the Journal of Policy, Regulation and Strategy for Telecommunications, Information and Media* 18(1). doi:10.1108/info-11-2015-0050.

Heywood, D., S. Cornelius & S. Carver. 2011. *An Introduction to Geographical Information Systems*. Fourth edi. London: Pearson Prentice Hall.

Hofmann, Charlotte. 2012. Representation and Perception of Mapped Space – The use of tourist maps in the imagining of the city. Uniersität Wien.

Hölzing, Jörg. 2008. Kano-Theorie der Kundenzufriedenheit. *Die Kano-Theorie der Kundenzufriedenheitsmessung: Eine theoretische und empirische Überprüfung*, 76–148. Wiesbaden: Gabler. doi:10.1007/978-3-8349-9864-4_3.

Hu, Yingjie. 2017. Geospatial Semantics. *Comprehensive Geographic Information Systems* 80–94. doi:10.1016/b978-0-12-409548-9.09597-x.

Hu, Yingjie. 2018. Geo-Text Data and Data-Driven Geospatial Semantics. *Geography Compass*. doi:10.1111/gec3.12404.

Hutchins, Ryan & Bill Mahoney. 2015. "Scraped" Airbnb data shows 58% of listings possibly illegal. *Politico*. https://www.politico.com/states/new-york/city-hall/story/2015/02/scraped-airbnb-data-shows-58-of-listings-possibly-illegal-019624.

Inkpen, Diana, Ji Liu, Atefeh Farzindar, Farzaneh Kazemi & Diman Ghazi. 2017. Location detection and disambiguation from twitter messages. *Journal of Intelligent*

Information Systems. Journal of Intelligent Information Systems 49(2). 237–253. doi:10.1007/s10844-017-0458-3.

Jackson, P & I Moulinier. 2002. *Natural Language Processing for Online Applications : Text Retrieval, Extraction, and Categorization*. Amsterdam: John Benjamins Publishing Co.

Jentsch, Heiko. 2009. Konzeption eines integrierten Qualitätsmanagements für den Stadtverkehr. TU Darmstadt.

Kainz, Wolfgang. 2004. The Mathematics of GIS.

Kainz, Wolfgang. 2007. Fuzzy Logic and GIS. *University of Vienna, Austria*.

Karagoz, Pinar, Halit Oguztuzun, Ruket Cakici, Ozer Ozdikis, Kezban Dilek Onal & Meryem Sagcan. 2016. Extracting Location Information from Crowd-sourced Social Network Data. *European Handbook of Crowdsourced Geographic Information*, 195–204. London: Ubiquity Press. doi:http://dx.doi.org/10.5334/bax.o.

Karimzadeh, Morteza, Scott Pezanowski, Alan M. MacEachren & Jan O. Wallgrün. 2019. GeoTxt: A scalable geoparsing system for unstructured text geolocation. *Transactions in GIS* 23(1). 118–136. doi:10.1111/tgis.12510.

Ke, Qing. 2017. Sharing Means Renting?: An Entire-Marketplace Analysis of Airbnb. *Proceedings of the 2017 ACM on Web Science Conference*. doi:10.2139/ssrn.2902840.

Khan, Arbaz, Maria Vasardani & Stephan Winter. 2013. Extracting Spatial Information From Place Descriptions. *ACM SIGSPATIAL COMP'13*, 62–69. Orlando, FL, USA. doi:10.1145/2534848.2534857.

Khanam, M. Humera, Md A. Khudhus & M. S.Prasad Babu. 2017. Named Entity Recognition using Machine learning techniques for Telugu language. *Proceedings of the IEEE International Conference on Software Engineering and Service Sciences, ICSESS*, 940–944. doi:10.1109/ICSESS.2016.7883220.

Kordopatis-Zilos, Giorgos, Symeon Papadopoulos & Ioannis Kompatsiaris. 2017. Geotagging Text Content with Language Models and Feature Mining. *Proceedings of the IEEE* 105(10). 1971–1986. doi:10.1109/JPROC.2017.2688799.

Kotzbek, Gilbert. 2016. GIS-gestützte Spielanalyse. Universität Wien.

Krämer, Hans. 1992. *Integrative Ethik*. Frankfurt am Main: Suhrkamp.

Krumm, John, Nigel Davies & Chandra Narayanaswami. 2008. User-Generated Content. *IEEE Pervasive Computing* 7(4). 10–11. doi:10.1109/MPRV.2008.85.

Kuhn, Werner. 2005. Geospatial Semantics: Why, of What, and How? In Stefano Spaccapietra & Esteban Zimányi (eds.), *Journal on Data Semantics III*, 1–24. Berlin, Heidelberg: Springer Berlin Heidelberg.

Lacity, Mary C. & Marius A. Janson. 1994. Understanding qualitative data: A framework of text analysis methods. *Journal of Management Information Systems* 11(2). 137–155. doi:10.1080/07421222.1994.11518043.

Lampoltshammer, Thomas J & Thomas Heistracher. 2012. Natural Language Processing in Geographic Information Systems-Some Trends and Open Issues. *International Journal of Computer Science and Emerging Technolgies* 3(3). 81–88.

Lee, Anthony J T, Ruey-Wen Hong, Wei-Min Ko, Wen-Kwang Tsao & Hsiu-Hui Lin. 2007. Mining spatial association rules in image databases. *Information Sciences* 177(7). 1593–1608. doi:https://doi.org/10.1016/j.ins.2006.09.018.

Leidner, Jochen L. & Michael D. Lieberman. 2011. Detecting geographical references in the form of place names and associated spatial natural language. *SIGSPATIAL Special* 3(2). 5–11. doi:10.1145/2047296.2047298.

Leonelli, Sabina. 2016. Locating ethics in data science: Responsibility and accountability in global and distributed knowledge production systems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374. doi:10.1098/rsta.2016.0122.

Lingad, John, Sarvnaz Karimi & Jie Yin. 2013. Location extraction from disaster-related microblogs. *Proceedings of the 22nd International Conference on World Wide Web - WWW '13 Companion*, 1017–1020. doi:10.1145/2487788.2488108.

Liu, Fei, Maria Vasardani & Timothy Baldwin. 2014. Automatic Identification of Locative Expressions from Social Media Text. *Proceedings of the 4th International Workshop on Location and the Web*, 9–16. doi:10.1145/2663713.2664426.

Ljunglöf, Peter & Mats Wiren. 2010. Syntactic Parsing. In Nitin Indurkhya & Fred J. Damerau (eds.), *Handbook of natural language processing*, 59–84. second edi. Boca Raton, Florida: CRC Press, Taylor & Francis Group.

Lloyd, Robert & Christopher Heivly. 1987. Systematic Distortions in Urban Cognitive Maps. *Annals of the Association of American Geographers*. John Wiley & Sons, Ltd (10.1111) 77(2). 191–207. doi:10.1111/j.1467-8306.1987.tb00153.x.

Lourenço, Rui Pedro. 2013. Open Government Portals Assessment: A Transparency for Accountability Perspective. In Maria A Wimmer, Marijn Janssen & Hans J Scholl (eds.), *Electronic Government*, 62–74. Springer Berlin Heidelberg.

Lukyanenko, Roman. 2014. An Information Modeling Approach to improve Quality of user-generated Content. *Dissertation*. Memorial University of Newfoundland.

Lynch, Kevin. 1960. *The Image of the City*. *MIT Press*. Cambridge and London. doi:10.2307/427643.

Ma, Xiao, Jeffery T. Hancock, Kenneth Lim Mingjie & Mor Naaman. 2017. Self-Disclosure and Perceived Trustworthiness of Airbnb Host Profiles. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW '17* 2397–2409. doi:10.1145/2998181.2998269.

Maggini, Marco. 2010. Natural language processing - Part 3: Syntax & grammar. Universität Siena. http://www.dii.unisi.it/~maggini/Teaching/TEL/slides EN/07 - NLP - Syntax.pdf (1 June, 2019).

Maogui, Hu & Wang Jinfeng. 2009. Application of automated testing tool in GIS modeling. *2009 WRI World Congress on Software Engineering, WCSE 2009* 4. 184–

188. doi:10.1109/WCSE.2009.94.

Mark, David & Andrew Frank. 1989. *Concepts of space and spatial language*.

Mark, David M & Max J Egenhofer. 1994. Modeling Spatial Relations Between Lines and Regions: Combining Formal Mathematical Models and Human Subjects Testing. *Cartography and Geographic Information Systems*. Taylor & Francis 21(4). 195–212. doi:10.1559/152304094782540637.

Martineau, Paris. 2019. Inside Airbnb's 'Guerrilla War' Against Local Governments. *Wired*. https://www.wired.com/story/inside-airbnbs-guerrilla-war-against-local-governments/.

Martinez, Angel R. 2010. Natural language processing. *Wiley Interdisciplinary Reviews: Computational Statistics* 2(3). 352–357. doi:10.1002/wics.76.

McKinney, Daene C & Ximing Cai. 2002. Linking GIS and water resources management models: an object-oriented method. *Environmental Modelling & Software* 17(5). 413–425. doi:https://doi.org/10.1016/S1364-8152(02)00015-4.

Middleton, Stuart E, Giorgos Kordopatis-Zilos, Symeon Papadopoulos & Yiannis Kompatsiaris. 2018. Location Extraction from Social Media : Geoparsing , Location Disambiguation and Geotagging. *ACM Transactions on Information Systems (TOIS)* 36(4). doi:10.1145/3202662.

Middleton, Stuart E & Vadims Krivcovs. 2015. Geoparsing and Geosemantics for Social Media: Spatiotemporal Grounding of Content Propagating Rumors to Support Trust and Veracity Analysis During Breaking News. *ACM Trans. Information Systems* 34(3). 1–27. doi:10.1145/2842604.

Mikheev, Andrei, Marc Moens & Claire Grover. 1999. Named Entity Recognition without Gazetteers. *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*. doi:10.3115/977035.977037.

Miner, Gary, Dursun Delen, John Elder, Andrew Fast, Thomas Hill & Robert A Nisbet (eds.). 2012. Chapter 2 - The Seven Practice Areas of Text Analytics. *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*, 29–41. Boston: Academic Press. doi:https://doi.org/10.1016/B978-0-12-386979-1.00002-5.

Moncla, Ludovic. 2016. Automatic reconstruction of itineraries from descriptive texts. Universität Saragossa.

Moncla, Ludovic, Mauro Gaio, Thierry Joliveau & Yves-françois Le Lay. 2017. Automated Geoparsing of Paris Street Names in 19th Century Novels. *1st ACM SIGSPATIAL Workshop on Geospatial Humanities*. doi:10.1145/3149858.3149859.

Moncla, Ludovic, Walter Renteria-agualimpia, Javier Nogueras-iso & Mauro Gaio. 2014. Geocoding for texts with fine-grain toponyms: an experiment on a geoparsed hiking descriptions corpus. *ACM SIGSPATIAL*. Dallas, Texas, United States.

Monica, N & KR Kumar. 2013. Survey on Big Data by Coordinating Mapreduce to Integrate Variety of Data. *International Journal of Science and Research* 2(11). 427–432.

Mutuku, Leonida N & Jessica Colaco. 2012. Increasing Kenyan Open Data Consumption:

A Design Thinking Approach. *Proceedings of the 6th International Conference on Theory and Practice of Electronic Governance* (ICEGOV '12), 18–21. New York, NY, USA: ACM. doi:10.1145/2463728.2463733.

Nadeau, David. 2007. Semi-Supervised Named Entity Recognition: Learning to Recognize 100 Entity Types with Little Supervision. University of Ottawa. doi:10.1075/li.30.1.03nad.

Nicolas, Moyroud & Portet Frédéric. 2018. Introduction to QGIS. *QGIS and Generic Tools*, 1–17. John Wiley & Sons, Ltd. doi:10.1002/9781119457091.ch1.

Palmer, David D. 2010. Text Preprocessing. In Nitin Indurkhya & Fred J. Damerau (eds.), *Handbook of natural language processing*, 9–28. second edi. Boca Raton, Florida: CRC Press, Taylor & Francis Group.

Palmer, David D. & David S. Day. 1997. A statistical profile of the Named Entity task. *ANLC '97 Proceedings of the fifth conference on Applied natural language processing*, 190–193. doi:10.3115/974557.974585.

Papadias, Dimitris & Timos Sellis. 1994. Qualitative Representation of Spatial Knowledge in Two-Dimensional Space. *VLDB Journal* 3. 479–516. doi:10.1007/bfb0020328.

Paterson, Laura L. & Ian Gregory. 2019. Geographical Information Systems and Textual Sources - Using Geographical Text Analysis to Understand Discourse. *Representations of Poverty and Place*, 41–60. Palgrave Macmillan.

Perumal, Mousi, Bhuvaneswari Velumani, Ananthi Sadhasivam & Kalpana Ramasami. 2015. Spatial Data Mining approaches for GIS – A brief review. *Advances in Intelligent Systems and Computing*, vol. 338. doi:10.1007/978-3-319-13731-5_63.

Petasis, Georgios, Frantz Vichot, Francis Wolinski, Georgios Paliouras, Vangelis Karkaletsis & Constantine D. Spyropoulos. 2001. Using machine learning to maintain rule-based named-entity recognition and classification systems. *ACL '01 Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, 426–433. doi:10.3115/1073012.1073067.

Peters, Dave. 2015. *The Evolution of GIS Software*.

Preethi, B Meena & P Radha. 2016. A Survey Paper on Text Mining - Techniques , Applications And Issues. *IOSR Journal of Computer Engineering (IOSR-JCE)* 46–51.

Purves, Ross S., Mark H. Hall, Paul Clough, Vanessa Murdock & Christopher B. Jones. 2018. Geographic Information Retrieval: Progress and Challenges in Spatial Search of Text. *Foundations and Trends® in Information Retrieval* 12(2–3). 164–318. doi:10.1561/1500000034.

Puschmann, Thomas & Rainer Alt. 2016. Sharing economy. *Business and Information Systems Engineering*. Springer Fachmedien Wiesbaden 58(1). 93–99. doi:10.1007/s12599-015-0420-2.

Reinhart, Gunther, Udo Lindemann & Joachim Heinzl. 1996. *Qualitätsmanagement: Ein Kurs für Studium und Praxis*.

Shahid, Rizwan, Stefania Bertazzon, Merril L. Knudtson & William A. Ghali. 2009. Comparison of distance measures in spatial analytical modeling for health service

planning. *BMC Health Services Research* 9. 1–14. doi:10.1186/1472-6963-9-200.

Sharma, J. 1996. Integrated Spatial Reasoning in Geographic Information Systems. University of Maine.

Shekhar, Shashi, Pusheng Zhang & Yan Huang. 2005. Spatial Data Mining. In Oded Maimon & Lior Rokach (eds.), *Data Mining and Knowledge Discovery Handbook*, 833–851. Boston, MA: Springer US. doi:10.1007/0-387-25465-X_39. https://doi.org/10.1007/0-387-25465-X_39.

Sint, Rolf, Sebastian Schaffert, Stephanie Stroka & Roland Ferstl. 2009. Combining unstructured, fully structured and semi-structured information in semantic wikis. *CEUR Workshop Proceedings* 464. 73–87.

Spranger, Michael & Simon Pauw. 2012. Dealing with Perceptual Deviation: Vague Semantics for Spatial Language and Quantification. In Luc Steels & Manfred Hild (eds.), *Language Grounding in Robots*, 173–192. Boston, MA: Springer US. doi:10.1007/978-1-4614-3064-3_9.

StackOverflow. 2019. Trends. *Stack Overflow Insights*. https://insights.stackoverflow.com/trends?tags=python (23 June, 2019).

Stadt Wien, MA23. 2016. *Vienna in Figures*.

Talen, Emily & Luc Anselin. 1998. Assessing Spatial Equity: An Evaluation of Measures of Accessibility to Public Playgrounds. *Environment and Planning A* 30. 595–613. doi:10.1068/a300595.

Talmy, Leonard. 1983. How Language Structures Space. In Herbert L Pick & Linda P Acredolo (eds.), *Spatial Orientation: Theory, Research, and Application*, 225–282. Boston, MA: Springer US. doi:10.1007/978-1-4615-9325-6_11.

Thurgar-Dawson, Chris. 2014. Reality Mining and Meaningful Motion Patterns: A Critical GIS for Literary Studies. In Ingo Berensmeyer & Christoph Ehland (eds.), *Perspectives on Mobility*, vol. 17, 69–88. Amsterdam, Netherlands: Rodopi.

Utomo, Muhammad Nur Yasir, Teguh Bharata Adji & Igi Ardiyanto. 2018. Geolocation prediction in social media data using text analysis: A review. *2018 International Conference on Information and Communications Technology (ICOIACT)* 84–89. doi:10.1109/ICOIACT.2018.8350674.

Vasardani, Maria, Sabine Timpf, Stephan Winter & Martin Tomko. 2013. From Descriptions to Depictions: A Conceptual Framework. In Thora Tenbrink, John Stell, Antony Galton & Zena Wood (eds.), *Spatial Information Theory*, 299–319. Cham: Springer International Publishing.

Wallgrün, Jan Oliver, Alexander Klippel & Timothy Baldwin. 2014. Building a corpus of spatial relational expressions extracted from web documents. *Proceedings of the 8th Workshop on Geographic Information Retrieval - GIR '14*, 1–8. doi:10.1145/2675354.2675702.

Walmsley, D. J. & J. M. Jenkins. 1992a. Cognitive Distance: A Neglected lssue in Travel Behavior. *Journal of Travel Research* 31(1). 24–29. doi:10.1177/004728759203100106.

Walmsley, D. J. & J. M. Jenkins. 1992b. Tourism cognitive mapping of unfamiliar environments. *Annals of Tourism Research* 19(2). 268–286. doi:10.1016/0160-7383(92)90081-Y.

Weibel, Robert. 2016. Analyse von Erreichbarkeit und Distanzbeziehungen. Universität Zürich.

Wickham, Hadley. 2019a. Tidyverse. *Overview*. https://tidyverse.tidyverse.org/ (26 June, 2019).

Wickham, Hadley. 2019b. Welcome to the tidyverse. https://speakerdeck.com/hadley/welcome-to-the-tidyverse (26 June, 2019).

Wilfinger, Lukas. 2018. Bildung einer Heuristik zur Plausibilisierung und Validierung von Verkehrsmanagementdaten. FH Wiener Neustadt.

Wilfling, Bernd. 2011. Folien zur Vorlesung Statistik I (Deskriptive Statistik). Westfälische Wilhelms-Universität Münster.

Wilke, Henry. 2013. Raum ist Zeit - Zeitentfernungskarten als Orientierungssystem im städtischen Fußverkehr. *Graue Reihe des Instituts für Stadt- und Regionalplanung*. TU Berlin.

Wissler, Lars, Mohammed Almashraee, Monett Dagmar & Adrian Paschke. 2014. The Gold Standard in Corpus Annotation. *Journal of Political Economy* 7(4). 551. doi:10.1086/250616.

Woodruff, Allison Gyle & Christian Plaunt. 1994. GIPSY : Automated Geographic Indexing of Text Documents. *Journal of the American Society for Information Science*.

Xu, Sen, Soner Flexner & Vitor Carvalho. 2012. Geocoding Billions of Addresses: Toward a Spatial Record Linkage System with Big Data. *GiBDA'12: Workshop on GIScience in the Big Data Age*, vol. 1, 17–26.

Yang, Tung-Mou, Jin Lo, Hui-Ju Wang & Jing Shiang. 2013. Open Data Development and Value-added Government Information: Case Studies of Taiwan e-Government. *Proceedings of the 7th International Conference on Theory and Practice of Electronic Governance* (ICEGOV '13), 238–241. New York, NY, USA: ACM. doi:10.1145/2591888.2591932.

Young, Martin. 1999. Cognitive maps of nature-based tourists. *Annals of Tourism Research* 26(4). 817–839. doi:10.1016/s0160-7383(99)00023-7.

Yuan, Yecheng. 2011. Extracting spatial relations from document for geographic information retrieval. *Proceedings - 2011 19th International Conference on Geoinformatics, Geoinformatics 2011*. doi:10.1109/GeoInformatics.2011.5980797.

Zhang, Wei & Judith Gelernter. 2014. Geocoding location expressions in Twitter messages: A preference learning method. *Journal of Spatial Information Science* 9(9). 37–70. doi:10.5311/JOSIS.2014.9.170.

Zhang, Ziqi. 2013. Named entity recognition: challenges in document annotation, gazetteer construction and disambiguation. The University of Sheffield.

Anhang I R-Code zur Extraktion von räumlichen und zeitlichen Distanzangaben aus den Bewertungen

```
####################################
# Laden der benötigten Pakete
####################################
library(dplyr)
library(tidytext)
library(readr)
library(stringr)

####################################
# Importierung der benötigten Dateien
####################################
# relativen Pfad zu den Datenordnern herstellen
setwd("C:/Users/Lukas
Wilfinger/Desktop/Masterarbeit_Uni_Wien/code_r/Production")
setwd("../../Daten")
datapath <- getwd()

# Import der Unterkünfte & Reviews
setwd(datapath)
setwd("Februar_Inside_airbnb")
listings_raw <- read.csv("listings.csv", encoding = "UTF-8",
stringsAsFactors = F)
reviews_raw <- read.csv("reviews.csv", encoding = "UTF-8",
stringsAsFactors = F)

# Import Lexika
setwd(datapath)
setwd("Gazetteer")
distance_lexika_time <- read.csv("lexikon_distanzen_zeit.txt", header =
T, encoding = "UTF-8", stringsAsFactors = F)
distance_lexika_space <- read.csv("lexikon_distanzen_raum.txt", header =
T, encoding = "UTF-8", stringsAsFactors = F)

##################
# Datenbereinigung
##################
# Filterung der relevanten Unterkünfte mittels der Unterkunft-IDs
filtered_listings <- listings_raw %>% filter(id %in% c(11581196, 358058,
14650908, 16144590, 24604164, 293667, 1915927, 16331770, 20989593,
1004037, 481895, 20842917, 5109180))

# Selektion der nur relevanten Spalten der Unterkünfte
selected_listings <- filtered_listings %>% select(id, name, summary,
space, description, transit, latitude, longitude)
rm(listings_raw)

# Join zwischen den selektierten Unterkünften und den dazugehörigen
Bewertungen
joined_reviews <-  selected_listings %>% left_join(reviews_raw, by =
c("id" = "listing_id"))
rm(reviews_raw, selected_listings, filtered_listings)

# Selektion der nur relevanten Spalten der Bewertungen
```

```
selected_reviews <- joined_reviews %>% select(id, comments)
###################
# Tokenization
###################
reviews_tokenized <- selected_reviews %>% unnest_tokens(word, comments)
# Selektion der numerischen Wörter = Zahlen
reviews_tokenized_numbers <-
suppressWarnings(parse_number(reviews_tokenized$word))

# Erstellen einer eindeutigen ID pro Zeile und Bewertung
reviews_ids <- rownames(reviews_tokenized)
# Zeilen ID den Bewertungen hinzufügen
reviews_ids_temp <- cbind(id_join=reviews_ids, reviews_tokenized)
# aus Zeilen ID und Zahlen einen Dataframe erstellen
reviews_id_temp_2 <- data.frame(reviews_ids_temp$id_join,
reviews_tokenized_numbers)

# Obigen Dataframe über die Zeilen ID mit den Bewertungen verknüpfen
# Es entsteht ein Dataframe, der Bewertungen, gefilterte Zahlen und
Zeilen ID enthält
temp_reviews <- reviews_ids_temp %>% left_join(reviews_id_temp_2, by =
c("id_join" = "reviews_ids_temp.id_join"))

###################
# Selektion der relevanten Wortfolgen nach 2 unterschiedlichen Mustern
# 1. Räumliche Distanzangaben:
#   -) 1. Muster der Zeichenfolge: XXXm oder XXXft
#   -) 2. Muster der Zeichenfolge: XXX m oder XXX ft (Beachte den
Abstand!)
# 2. Zeitliche Distanzangaben:
#   -) 1. Muster der Zeichenfolge: XXXmin oder XXXMinuten
#   -) 2. Muster der Zeichenfolge: XXX min oder XXX Minuten (Beachte den
Abstand!)
###################
# Funktion mit der, die umgebenden Wörter, bei Auftreten eines der
obigen Muster, selektiert werden
# Quelle: https://stackoverflow.com/questions/13155609/returning-above-
and-below-rows-of-specific-rows-in-r-dataframe
(aufgerufen am: 11.03.2019)

getMyRows <- function(data, matches, range) {
  rowMatches = lapply(unlist(lapply(matches, function(x)
    grep(x, rownames(data)))), function(y) y + range)
  rowMatches = lapply(rowMatches, function(x) x[x > 0 & x <=
nrow(data)])
  lapply(rowMatches, function(x) data[x, ])
}
```

```
##############################
# Räumliche Distanzen: 1.Muster
##############################
# Erzeugen eines leeren Dataframes, der die gleiche Länge, wie der
Bewertungs-Dataframe hat
df_reviews <- as.data.frame(matrix(0, nrow(temp_reviews), 1))

# Extraktion des letzten Buchstaben eines Wortes
for (i in 1:nrow(temp_reviews)){
  df_reviews[i,] <- substr(temp_reviews$word[i],
str_length(temp_reviews$word[i]), str_length(temp_reviews$word[i]))
}

# Hinzufügen der extrahierten Buchstaben zu den Zeilen IDs
reviews_comment_ending <- cbind(df_reviews, reviews_ids)

# Verknüpfung der einzelnen Buchstaben und der tokenizierten
Bewertungstabelle
# Annahme: Alle Angaben erfolgen in Metern (m) oder Feet (ft ==> t)
reviews_distances_spatial <- temp_reviews %>%
left_join(reviews_comment_ending, by = c("id_join" = "reviews_ids")) %>%
  filter(is.na(reviews_tokenized_numbers) == F &
           reviews_tokenized_numbers != V1 &
           word != reviews_tokenized_numbers &
           (V1 == "m" | V1 == "t"))

# Den herausgefilterten Zeilen wird eine Flag zugeordnet: spatial_found
= Yes
reviews_distances_spatial <- reviews_distances_spatial %>%
mutate(spatial_found = "Yes")

# Verknüpfung der geflagten Zeilen IDs und der tokenizierten
Bewertungstabelle
reviews_spatial_temp <- temp_reviews %>%
left_join(reviews_distances_spatial, by = "id_join")

# Erstellung von Zeilennamen
rownames(reviews_spatial_temp) <- paste("rowname",
reviews_spatial_temp$id_join,

reviews_spatial_temp$spatial_found, sep = ".")

# Selektion der 12 vorherigen und 12 nachgereihten Wörter, bei einer
räumlichen Angabe
# Wenn Zeilenname "Yes" enthält, dann selektierte die vorherigen und
nachfolgenden 12 Wörter
spatial_temp <- getMyRows(reviews_spatial_temp, c("Yes"), -12:12)
# Aus der entstehenden Liste an Dataframes muss ein großer Dataframe
entstehen
spatial_temp_1 <- bind_rows(spatial_temp, .id = "id_join")
# Umbenennung der ID-Spalte
colnames(spatial_temp_1)[1] <- "DF_ID_listings_spatial"
# Kennzeichnung, um welche Art der Distanzangabe es sich handelt
spatial_temp_1 <- spatial_temp_1  %>% mutate(type_of_recognition =
"Spatial")
# Selektion nur der relevanten Spalten
spatial_temp_1  <- spatial_temp_1  %>% select("id.x", "id_join",
"word.x","type_of_recognition")
# Umbennenung weiterer Spalten
```

```
colnames(spatial_temp_1 )[1] <- "id"
colnames(spatial_temp_1 )[3] <- "word"

##############################
# Räumliche Distanzen: 2.Muster
##############################
# Finden von räumlichen Distanzangaben in einem eigens erstellten
Lexikon: Beispiele: m, mts, ft, mi, Meter, etc.
reviews_temp_space <- temp_reviews %>%
  left_join(distance_lexika_space, by = c("word" =
"Lexikon_Distanz_Raum"))
# Hinzufügen, dass es sich um eine räumliche Angabe handelt
reviews_temp_space <- reviews_temp_space %>%
  mutate(info_space = "Raeumliche Angabe")
# Herausfiltern der vorherigen Zeile, auf die sich die Distanzangabe
bezieht
reviews_temp_space <- reviews_temp_space %>%
  mutate(string_part_space = ifelse(lead(Flag==1, n = 1), info_space,
0))

# Erstellen von Zeilennamen
rownames(reviews_temp_space) <- paste("rowname",
reviews_temp_space$id_join,

reviews_temp_space$string_part_space, sep = ".")

# Suchen nach dem Muster "Räumliche Angaben" und Selektion der
vorherigen und nachfolgenden 12 Wörter
reviews_test_space <- getMyRows(reviews_temp_space, c("Raeumliche
Angabe"), -12:12)

# Aus einer Liste aus Dataframes, einen erzeugen
reviews_relevant_space <- bind_rows(reviews_test_space, .id = "id_join")
# Spaltenname ändern
colnames(reviews_relevant_space )[1] <- "DF_ID_reviews_space"
# Spalte hinzufügen
reviews_relevant_space  <- reviews_relevant_space  %>%
  mutate(type_of_recognition = "space")
# Nur relevante Spalten selektieren
reviews_relevant_space  <- reviews_relevant_space  %>%
  select("id", "id_join", "word", "type_of_recognition",
"reviews_tokenized_numbers", "string_part_space")

# Zeilen IDs selektieren, die numerisch sind und einen Raumbezug
aufweisen
reviews_ids_temp_space <- reviews_relevant_space %>%
  filter(reviews_tokenized_numbers != 'NA' & is.na(string_part_space) ==
F) %>%
  select(id)

# Zeilen IDs als character casten
reviews_relevant_space_temp <-
as.character(reviews_relevant_space$id_join)
# Teilen der Zeilen IDs -> dadurch ist später eine Zuordnung zu
zusammengehörige Sätze möglich
reviews_relevant_space_temp <- strsplit(reviews_relevant_space_temp,
split ='.', fixed=TRUE)
# Maximale Länge/Breite der Liste bestimmen
len <- max(lengths(reviews_relevant_space_temp))
```

```
# Wenn Liste kürzer als maximale Länge, dann mit NAs füllen
cols <- lapply(reviews_relevant_space_temp, function(l) c(l, rep(NA, len
- length(l))))
# Liste in einen Dataframe umwandeln
reviews_relevant_space_temp <- data.frame(matrix(unlist(cols),
nrow=length(cols), byrow=T))

reviews_relevant_space_temp <- cbind(reviews_relevant_space,
reviews_relevant_space_temp)

reviews_ids_temp_space <- reviews_relevant_space_temp %>%
  filter(reviews_tokenized_numbers != 'NA' & is.na(string_part_space) ==
F) %>%
  select(X1)

# Nur Sätze selektieren, bei denen ein räumlich relevanter Kontext mit
Zahlen besteht
reviews_relevant_space_temp <- reviews_relevant_space_temp %>%
inner_join(reviews_ids_temp_space, by = "X1") %>%
  distinct(id, id_join, word, type_of_recognition,
'reviews_tokenized_numbers', 'string_part_space')
reviews_relevant_space_temp$type_of_recognition <- "Spatial"
# Finalen Dataframe dieses Musters bilden
spatial_temp_2 <- reviews_relevant_space_temp %>% select("id",
"id_join", "word", "type_of_recognition")

# Beide erstellten Dataframes in einen verbinden
reviews_spatial_final <- rbind(spatial_temp_1, spatial_temp_2)

################################
# Zeitliche Distanzen: 1.Muster
################################

df_reviews <- as.data.frame(matrix(0, nrow(temp_reviews), 1))

# Extraktion des letzten Buchstaben eines Wortes
for (i in 1:nrow(temp_reviews)){
  df_reviews[i,] <- substr(temp_reviews$word[i],
str_length(temp_reviews$word[i]), str_length(temp_reviews$word[i]))
}

# Hinzufügen der extrahierten Buchstaben zu den Zeilen IDs
reviews_comment_ending <- cbind(df_reviews, reviews_ids)

# Verknüpfung der einzelnen Buchstaben und der tokenizierten
Bewertungstabelle
# Annahme: Alle Angaben erfolgen in Metern (m) oder Feet (ft ==> t)
reviews_distances_temporal <- temp_reviews %>%
left_join(reviews_comment_ending, by = c("id_join" = "reviews_ids")) %>%
  filter(is.na(reviews_tokenized_numbers) == F &
           reviews_tokenized_numbers != V1 &
           word != reviews_tokenized_numbers &
           (V1 == "n"))

# Den herausgefilterten Zeilen wird eine Flag zugeordnet: temporal_found
= Yes
reviews_distances_temporal <- reviews_distances_temporal %>%
mutate(temporal_found = "Yes")
```

```
# Verknüpfung der geflagten Zeilen IDs und der tokenizierten
Bewertungstabelle
reviews_temporal_temp <- temp_reviews %>%
left_join(reviews_distances_temporal, by = "id_join")

# Erstellung von Zeilennamen
rownames(reviews_temporal_temp) <- paste("rowname",
reviews_temporal_temp$id_join,

reviews_temporal_temp$temporal_found, sep = ".")

# Selektion der 12 vorherigen und 12 nachgereihten Wörter, bei einer
räumlichen Angabe
# Wenn Zeilenname "Yes" enthält, dann selektierte die vorherigen und
nachfolgenden 12 Wörter
temporal_temp <- getMyRows(reviews_temporal_temp, c("Yes"), -12:12)
# Aus der entstehenden Liste an Dataframes muss ein großer Dataframe
entstehen
temporal_temp_1 <- bind_rows(temporal_temp, .id = "id_join")
# Umbenennung der ID-Spalte
colnames(temporal_temp_1)[1] <- "DF_ID_listings_temporal"
# Kennzeichnung, um welche Art der Distanzangabe es sich handelt
temporal_temp_1 <- temporal_temp_1  %>% mutate(type_of_recognition =
"Temporal")
# Selektion nur der relevanten Spalten
temporal_temp_1  <- temporal_temp_1  %>% select("id.x", "id_join",
"word.x","type_of_recognition")
# Umbennenung weiterer Spalten
colnames(temporal_temp_1 )[1] <- "id"
colnames(temporal_temp_1 )[3] <- "word"

################################
# Zeitliche Distanzen: 2.Muster
################################

reviews_temp_temporal <- temp_reviews %>%
  left_join(distance_lexika_time, by = c("word" =
"Lexikon_Distanz_Zeit"))
# Hinzufügen, dass es sich um eine zeitliche Angabe handelt
reviews_temp_temporal <- reviews_temp_temporal %>%
  mutate(info_temporal = "Zeitliche Angabe")
# Herausfiltern der vorherigen Zeile, auf die sich die Distanzangabe
bezieht
reviews_temp_temporal <- reviews_temp_temporal %>%
  mutate(string_part_temporal = ifelse(lead(Flag==1, n = 1),
info_temporal, 0))

# Erstellen von Zeilennamen
rownames(reviews_temp_temporal) <- paste("rowname",
reviews_temp_temporal$id_join,

reviews_temp_temporal$string_part_temporal, sep = ".")

# Suchen nach dem Muster "Zeitliche Angaben" und Selektion der
vorherigen und nachfolgenden 12 Wörter
reviews_test_temporal <- getMyRows(reviews_temp_temporal, c("Zeitliche
Angabe"), -12:12)

# Aus einer Liste an Dataframes, einen erzeugen
```

```
reviews_relevant_temporal <- bind_rows(reviews_test_temporal, .id =
"id_join")
# Spaltenname ändern
colnames(reviews_relevant_temporal)[1] <- "DF_ID_reviews_temporal"
# Spalte hinzufügen
reviews_relevant_temporal  <- reviews_relevant_temporal  %>%
  mutate(type_of_recognition = "Temporal")
# Nur relevante Spalten selektieren
reviews_relevant_temporal  <- reviews_relevant_temporal  %>%
  select("id", "id_join", "word", "type_of_recognition",
"reviews_tokenized_numbers", "string_part_temporal")

# Zeilen IDs selektieren, die numerisch sind und einen Raumbezug
aufweisen
reviews_ids_temp_temporal <- reviews_relevant_temporal %>%
  filter(reviews_tokenized_numbers != 'NA' & is.na(string_part_temporal)
== F) %>%
  select(id)

# Zeilen IDs als character casten
reviews_relevant_temporal_temp <-
as.character(reviews_relevant_temporal$id_join)
# Teilen der Zeilen IDs -> dadurch ist später eine Zuordnung zu
zusammengehörige Sätze möglich
reviews_relevant_temporal_temp <-
strsplit(reviews_relevant_temporal_temp, split ='.', fixed=TRUE)
# Maximale Länge/Breite der Liste bestimmen
len <- max(lengths(reviews_relevant_temporal_temp))
# Wenn Liste kürzer als maximale Länge, dann mit NAs füllen
cols <- lapply(reviews_relevant_temporal_temp, function(l) c(l, rep(NA,
len - length(l))))
# Liste in einen Dataframe umwandeln
reviews_relevant_temporal_temp <- data.frame(matrix(unlist(cols),
nrow=length(cols), byrow=T))

reviews_relevant_temporal_temp <- cbind(reviews_relevant_temporal,
reviews_relevant_temporal_temp)

reviews_ids_temp_temporal <- reviews_relevant_temporal_temp %>%
  filter(reviews_tokenized_numbers != 'NA' & is.na(string_part_temporal)
== F) %>%
  select(X1)

# Nur Sätze selektieren, bei denen ein räumlich relevanter Kontext mit
Zahlen besteht
reviews_relevant_temporal_temp <- reviews_relevant_temporal_temp %>%
inner_join(reviews_ids_temp_temporal, by = "X1") %>%
  distinct(id, id_join, word, type_of_recognition,
'reviews_tokenized_numbers', 'string_part_temporal')
reviews_relevant_temporal_temp$type_of_recognition <- "Temporal"
# Finalen Dataframe dieses Musters bilden
temporal_temp_2 <- reviews_relevant_temporal_temp %>% select("id",
"id_join", "word", "type_of_recognition")

# Beide erstellten Dataframes in einen verbinden
reviews_temporal_final <- rbind(temporal_temp_1, temporal_temp_2)
reviews_final <- rbind(reviews_spatial_final, reviews_temporal_final)
reviews_final <- reviews_final %>% mutate(Source = "Reviews")
```

```
#####################################
# Pivotieren der Wörter in Satzgruppen
#####################################
# Jede Bewertung enthält eine eindeutige ID
# Nach dieser ID sollen die einzelnen Wörter zu einem Satz (ohne
Satzzeichen)
# zusammengefügt werden
# Der Code ist, bis auf die Variablennamen gleich wie weiter oben
# Am Ende wird der finale Dataframe als CSV exportiert und händisch
codiert
reviews_pivot_id_1 <- as.character(reviews_final$id_join)
# Teilen der Zeilen IDs -> dadurch ist später eine Zuordnung zu
zusammengehörige Sätze möglich
reviews_pivot_id_1 <- strsplit(reviews_pivot_id_1, split ='.',
fixed=TRUE)
# Maximale Länge/Breite der Liste bestimmen
len <- max(lengths(reviews_pivot_id_1))
# Wenn Liste kürzer als maximale Länge, dann mit NAs füllen
reviews_pivot_id_1_cols <- lapply(reviews_pivot_id_1, function(l) c(l,
rep(NA, len - length(l))))
# Liste in einen Dataframe umwandeln
reviews_pivot_id_temp <-
data.frame(matrix(unlist(reviews_pivot_id_1_cols),
nrow=length(reviews_pivot_id_1_cols), byrow=T))

reviews_final_pivot_temp <- cbind(reviews_final, reviews_pivot_id_temp)

reviews_final_pivot <- reviews_final_pivot_temp %>%
  group_by(id, type_of_recognition, Source, X1) %>%
  summarise(sätze = paste(word, collapse=" "))
```

Anhang II R-Code zur Extraktion von räumlichen und zeitlichen Distanzangaben aus den Beschreibungen

```
####################################
# Laden der benötigten Pakete
####################################
library(dplyr)
library(tidytext)
library(readr)
library(stringr)

####################################
# Importierung der benötigten Dateien
####################################
# relativen Pfad zu den Datenordnern herstellen
setwd("C:/Users/Lukas
Wilfinger/Desktop/Masterarbeit_Uni_Wien/code_r/Production")
setwd("../../Daten")
datapath <- getwd()

# Import der Unterkünfte & Reviews
setwd(datapath)
setwd("Februar_Inside_airbnb")
listings_raw <- read.csv("listings.csv", encoding = "UTF-8",
stringsAsFactors = F)
reviews_raw <- read.csv("reviews.csv", encoding = "UTF-8",
stringsAsFactors = F)

# Import Lexika
setwd(datapath)
setwd("Gazetteer")
distance_lexika_time <- read.csv("lexikon_distanzen_zeit.txt", header =
T, encoding = "UTF-8", stringsAsFactors = F)
distance_lexika_space <- read.csv("lexikon_distanzen_raum.txt", header =
T, encoding = "UTF-8", stringsAsFactors = F)

##################
# Datenbereinigung
##################
# Filterung der relevanten Unterkünfte mittels der Unterkunft-IDs
filtered_listings <- listings_raw %>% filter(id %in% c(11581196, 358058,
14650908, 16144590, 24604164, 293667, 1915927, 16331770, 20989593,
1004037, 481895, 20842917, 5109180))

# Selektion der nur relevanten Spalten der Unterkünfte
selected_listings <- filtered_listings %>% select(id, name, summary,
space, description, transit, latitude, longitude)
rm(listings_raw)

##################
# Tokenization
##################
listings_tokenized <- selected_listings %>% unnest_tokens(word, transit)
# Selektion der numerischen Wörter = Zahlen
listings_tokenized_numbers <-
suppressWarnings(parse_number(listings_tokenized$word))
```

```
# Erstellen einer eindeutigen ID pro Zeile und Bewertung
listings_ids <- rownames(listings_tokenized)
# Zeilen ID den Bewertungen hinzufügen
listings_ids_temp <- cbind(id_join=listings_ids, listings_tokenized)
# aus Zeilen ID und Zahlen einen Dataframe erstellen
listings_id_temp_2 <- data.frame(listings_ids_temp$id_join,
listings_tokenized_numbers)

# Obigen Dataframe über die Zeilen ID mit den Bewertungen verknüpfen
# Es entsteht ein Dataframe, der Bewertungen, gefilterte Zahlen und
Zeilen ID enthält
temp_listings <- listings_ids_temp %>% left_join(listings_id_temp_2, by
= c("id_join" = "listings_ids_temp.id_join"))

##################
# Selektion der relevanten Wortfolgen nach 2 unterschiedlichen Mustern
# 1. Räumliche Distanzangaben:
#    -) 1. Muster der Zeichenfolge: XXXm oder XXXft
#    -) 2. Muster der Zeichenfolge: XXX m oder XXX ft (Beachte den
Abstand!)
# 2. Zeiliche Distanzangaben:
#    -) 1. Muster der Zeichenfolge: XXXmin oder XXXMinuten
#    -) 2. Muster der Zeichenfolge: XXX min oder XXX Minuten (Beachte den
Abstand!)
##################

# Funktion mit der, die umgebenden Wörter, bei auftreten eines der
obigen Muster, selektiert werden
# Quelle: https://stackoverflow.com/questions/13155609/returning-above-
and-below-rows-of-specific-rows-in-r-dataframe
#(aufgerufen am: 11.03.2019)
getMyRows <- function(data, matches, range) {
  rowMatches = lapply(unlist(lapply(matches, function(x)
    grep(x, rownames(data)))), function(y) y + range)
  rowMatches = lapply(rowMatches, function(x) x[x > 0 & x <=
nrow(data)])
  lapply(rowMatches, function(x) data[x, ])
}

##############################
# Räumliche Distanzen: 1.Muster
##############################
# Erzeugen eines leeren Dataframes, der die gleiche Länge, wie der
Bewertungs-Dataframe hat
df_listings <- as.data.frame(matrix(0, nrow(temp_listings), 1))

# Extraktion des letzten Buchstaben eines Wortes
for (i in 1:nrow(temp_listings)){
  df_listings[i,] <- substr(temp_listings$word[i],
str_length(temp_listings$word[i]), str_length(temp_listings$word[i]))
}

# Hinzufügen der extrahierten Buchstaben zu den Zeilen IDs
listings_comment_ending <- cbind(df_listings, listings_ids)
```

```
# Verknüpfung der einzelnen Buchstaben und der tokenizierten
Bewertungstabelle
# Annahme: Alle Angaben erfolgen in Metern (m) oder Feet (ft ==> t)
listings_distances_spatial <- temp_listings %>%
left_join(listings_comment_ending, by = c("id_join" = "listings_ids"))
%>%
  filter(is.na(listings_tokenized_numbers) == F &
           listings_tokenized_numbers != V1 &
           word != listings_tokenized_numbers &
           (V1 == "m" | V1 == "t"))

# Den herausgefilterten Zeilen wird eine Flag zugeordnet: spatial_found
= Yes
listings_distances_spatial <- listings_distances_spatial %>%
mutate(spatial_found = "Yes")

# Verknüpfung der geflagten Zeilen IDs und der tokenizierten
Bewertungstabelle
listings_spatial_temp <- temp_listings %>%
left_join(listings_distances_spatial, by = "id_join")

# Erstellung von Zeilennamen
rownames(listings_spatial_temp) <- paste("rowname",
listings_spatial_temp$id_join,

listings_spatial_temp$spatial_found, sep = ".")

# Selektion der 12 vorherigen und 12 nachgereihten Wörter, bei einer
räumlichen Angabe
# Wenn Zeilenname "Yes" enthält, dann selektierte die vorherigen und
nachfolgenden 12 Wörter
spatial_temp <- getMyRows(listings_spatial_temp, c("Yes"), -12:12)
# Aus der entstehenden Liste an Dataframes muss ein großer Dataframe
entstehen
spatial_temp_listings_1 <- bind_rows(spatial_temp, .id = "id_join")
# Umbenennung der ID-Spalte
colnames(spatial_temp_listings_1)[1] <- "DF_ID_listings_spatial"
# Kennzeichnung, um welche Art der Distanzangabe es sich handelt
spatial_temp_listings_1 <- spatial_temp_listings_1  %>%
mutate(type_of_recognition = "Spatial")
# Selektion nur der relevanten Spalten
spatial_temp_listings_1  <- spatial_temp_listings_1  %>% select("id.x",
"id_join", "word.x","type_of_recognition")
# Umbennenung weiterer Spalten
colnames(spatial_temp_listings_1 )[1] <- "id"
colnames(spatial_temp_listings_1 )[3] <- "word"

################################
# Räumliche Distanzen: 2.Muster
################################
# Finden von räumlichen Distanzangaben in einem eigens erstellten
Lexikon: Beispiele: m, mts, ft, mi, Meter, etc.
listings_temp_space <- temp_listings %>%
  left_join(distance_lexika_space, by = c("word" =
"Lexikon_Distanz_Raum"))
# Hinzufügen, dass es sich um eine räumliche Angabe handelt
listings_temp_space <- listings_temp_space %>%
  mutate(info_space = "Raeumliche Angabe")
```

```
# Herausfiltern der vorherigen Zeile, auf die sich die Distanzangabe
bezieht
listings_temp_space <- listings_temp_space %>%
  mutate(string_part_space = ifelse(lead(Flag==1, n = 1), info_space,
0))

# Erstellen von Zeilennamen
if(length(is.na(listings_temp_space$string_part_space)) !=
length(listings_temp_space$string_part_space)){
rownames(listings_temp_space) <- paste("rowname",
listings_temp_space$id_join,

listings_temp_space$string_part_space, sep = ".")


# Suchen nach dem Muster "Räumliche Angaben" und Selektion der
vorherigen und nachfolgenden 12 Wörter

listings_test_space <- getMyRows(listings_temp_space, c("Raeumliche
Angabe"), -12:12)

# Aus einer Liste aus Dataframes, einen erzeugen
listings_relevant_space <- bind_rows(listings_test_space, .id =
"id_join")
# Spaltenname ändern
colnames(listings_relevant_space )[1] <- "DF_ID_listings_space"
# Spalte hinzufügen
listings_relevant_space  <- listings_relevant_space  %>%
  mutate(type_of_recognition = "space")
# Nur relevante Spalten selektieren
listings_relevant_space  <- listings_relevant_space  %>%
  select("id", "id_join", "word", "type_of_recognition",
"listings_tokenized_numbers", "string_part_space")

# Zeilen IDs selektieren, die numerisch sind und einen Raumbezug
aufweisen
listings_ids_temp_space <- listings_relevant_space %>%
  filter(listings_tokenized_numbers != 'NA' & is.na(string_part_space)
== F) %>%
  select(id)

# Zeilen IDs als character casten
listings_relevant_space_temp <-
as.character(listings_relevant_space$id_join)
# Teilen der Zeilen IDs -> dadurch ist später eine Zuordnung zu
zusammengehörigen Sätzen möglich
listings_relevant_space_temp <- strsplit(listings_relevant_space_temp,
split ='.', fixed=TRUE)
# Maximale Länge/Breite der Liste bestimmen
len <- max(lengths(listings_relevant_space_temp))
# Wenn Liste kürzer als maximale Länge, dann mit NAs füllen
cols <- lapply(listings_relevant_space_temp, function(l) c(l, rep(NA,
len - length(l))))
# Liste in einen Dataframe umwandeln
listings_relevant_space_temp <- data.frame(matrix(unlist(cols),
nrow=length(cols), byrow=T))

listings_relevant_space_temp <- cbind(listings_relevant_space,
listings_relevant_space_temp)
```

```
listings_ids_temp_space <- listings_relevant_space_temp %>%
  filter(listings_tokenized_numbers != 'NA' & is.na(string_part_space)
== F) %>%
  select(X1)

# Nur Sätze selektieren, bei denen ein räumlich relevanter Kontext mit
Zahlen besteht
listings_relevant_space_temp <- listings_relevant_space_temp %>%
inner_join(listings_ids_temp_space, by = "X1") %>%
  distinct(id, id_join, word, 'Type of Recoginition',
'listings_tokenized_numbers', 'string_part_space')
listings_relevant_space_temp$type_of_recognition <- "Spatial"

# Finalen Dataframe dieses Musters bilden
spatial_temp_2 <- listings_relevant_space_temp %>% select("id",
"id_join", "word", "type_of_recognition")
listings_spatial_final <- rbind(spatial_temp_1, spatial_temp_2)
} else{
  print("Keine Informationen mit räumlichen Muster 2 vorhanden")
# Beide erstellten Dataframes in einen verbinden
listings_spatial_final <- spatial_temp_listings_1
}

################################
# Zeitliche Distanzen: 1.Muster
################################

df_listings <- as.data.frame(matrix(0, nrow(temp_listings), 1))

# Extraktion des letzten Buchstaben eines Wortes
for (i in 1:nrow(temp_listings)){
  df_listings[i,] <- substr(temp_listings$word[i],
str_length(temp_listings$word[i]), str_length(temp_listings$word[i]))
}

# Hinzufügen der extrahierten Buchstaben zu den Zeilen IDs
listings_comment_ending <- cbind(df_listings, listings_ids)

# Verknüpfung der einzelnen Buchstaben und der tokenizierten
Bewertungstabelle
# Annahme: Alle Angaben erfolgen in Minuten (n)
listings_distances_temporal <- temp_listings %>%
left_join(listings_comment_ending, by = c("id_join" = "listings_ids"))
%>%
  filter(is.na(listings_tokenized_numbers) == F &
           listings_tokenized_numbers != V1 &
           word != listings_tokenized_numbers &
           (V1 == "n"))

# Den herausgefilterten Zeilen wird eine Flag zugeordnet: temporal_found
= Yes
listings_distances_temporal <- listings_distances_temporal %>%
mutate(temporal_found = "Yes")

# Verknüpfung der geflagten Zeilen IDs und der tokenizierten
Bewertungstabelle
listings_temporal_temp <- temp_listings %>%
left_join(listings_distances_temporal, by = "id_join")
```

```
# Erstellung von Zeilennamen
if(length(is.na(listings_temporal_temp$temporal_found)) !=
length(listings_temporal_temp$temporal_found)){
rownames(listings_temporal_temp) <- paste("rowname",
listings_temporal_temp$id_join,

listings_temporal_temp$temporal_found, sep = ".")

# Selektion der 12 vorherigen und 12 nachgereihten Wörter, bei einer
räumlichen Angabe
# Wenn Zeilenname "Yes" enthält, dann selektierte die vorherigen und
nachfolgenden 12 Wörter
temporal_temp <- getMyRows(listings_temporal_temp, c("Yes"), -12:12)
# Aus der entstehenden Liste an Dataframes muss ein großer Dataframe
entstehen
temporal_temp_listings_1 <- bind_rows(temporal_temp, .id = "id_join")
# Umbenennung der ID-Spalte
colnames(temporal_temp_listings_1)[1] <- "DF_ID_listings_temporal"
# Kennzeichnung, um welche Art der Distanzangabe es sich handelt
temporal_temp_listings_1 <- temporal_temp_listings_1  %>%
mutate(type_of_recognition = "Temporal")
# Selektion nur der relevanten Spalten
temporal_temp_listings_1  <- temporal_temp_listings_1  %>%
select("id.x", "id_join", "word.x","type_of_recognition")
# Umbennenung weiterer Spalten
colnames(temporal_temp_listings_1 )[1] <- "id"
colnames(temporal_temp_listings_1 )[3] <- "word"
}else{
  print("Keine Informationen mit zeitlichen Muster 1 vorhanden")
}

################################
# Zeitliche Distanzen: 2.Muster
################################

listings_temp_temporal <- temp_listings %>%
  left_join(distance_lexika_time, by = c("word" =
"Lexikon_Distanz_Zeit"))
# Hinzufügen, dass es sich um eine zeitliche Angabe handelt
listings_temp_temporal <- listings_temp_temporal %>%
  mutate(info_temporal = "Zeitliche Angabe")
# Herausfiltern der vorherigen Zeile, auf die sich die Distanzangabe
bezieht
listings_temp_temporal <- listings_temp_temporal %>%
  mutate(string_part_temporal = ifelse(lead(Flag==1, n = 1),
info_temporal, 0))

# Erstellen von Zeilennamen
rownames(listings_temp_temporal) <- paste("rowname",
listings_temp_temporal$id_join,

listings_temp_temporal$string_part_temporal, sep = ".")

# Suchen nach dem Muster "Zeitliche Angaben" und Selektion der
vorherigen und nachfolgenden 12 Wörter
listings_test_temporal <- getMyRows(listings_temp_temporal, c("Zeitliche
Angabe"), -12:12)
```

```
# Aus einer Liste an Dataframes, einen erzeugen
listings_relevant_temporal <- bind_rows(listings_test_temporal, .id =
"id_join")
# Spaltenname ändern
colnames(listings_relevant_temporal)[1] <- "DF_ID_listings_temporal"
# Spalte hinzufügen
listings_relevant_temporal  <- listings_relevant_temporal  %>%
  mutate(type_of_recognition = "Temporal")
# Nur relevante Spalten selektieren
listings_relevant_temporal  <- listings_relevant_temporal  %>%
  select("id", "id_join", "word", "type_of_recognition",
"listings_tokenized_numbers", "string_part_temporal")
# Zeilen IDs selektieren, die numerisch sind und einen Raumbezug
aufweisen
listings_ids_temp_temporal <- listings_relevant_temporal %>%
  filter(listings_tokenized_numbers != 'NA' &
is.na(string_part_temporal) == F) %>%
  select(id)

# Zeilen IDs als character casten
listings_relevant_temporal_temp <-
as.character(listings_relevant_temporal$id_join)
# Teilen der Zeilen IDs -> dadurch ist später eine Zuordnung zu
zusammengehörigen Sätzen möglich
listings_relevant_temporal_temp <-
strsplit(listings_relevant_temporal_temp, split ='.', fixed=TRUE)
# Maximale Länge/Breite der Liste bestimmen
len <- max(lengths(listings_relevant_temporal_temp))
# Wenn Liste kürzer als maximale Länge, dann mit NAs füllen
cols <- lapply(listings_relevant_temporal_temp, function(l) c(l, rep(NA,
len - length(l))))
# Liste in einen Dataframe umwandeln
listings_relevant_temporal_temp <- data.frame(matrix(unlist(cols),
nrow=length(cols), byrow=T))

listings_relevant_temporal_temp <- cbind(listings_relevant_temporal,
listings_relevant_temporal_temp)

listings_ids_temp_temporal <- listings_relevant_temporal_temp %>%
  filter(listings_tokenized_numbers != 'NA' &
is.na(string_part_temporal) == F) %>%
  select(X1)

# Nur Sätze selektieren, bei denen ein räumlich relevanter Kontext mit
Zahlen besteht
listings_relevant_temporal_temp <- listings_relevant_temporal_temp %>%
inner_join(listings_ids_temp_temporal, by = "X1") %>%
  distinct(id, id_join, word, type_of_recognition,
'listings_tokenized_numbers', 'string_part_temporal')
listings_relevant_temporal_temp$type_of_recognition <- "Temporal"
# Finalen Dataframe dieses Musters bilden
temporal_temp_listings_2 <- listings_relevant_temporal_temp %>%
select("id", "id_join", "word", "type_of_recognition")

# Beide erstellten Dataframes in einen verbinden
if(exists("temporal_temp_listings_1") == TRUE){
listings_temporal_final <- rbind(temporal_temp_listings_1,
temporal_temp_listings_2)
}else{print("Nur 2.zeitliches Muster nicht leer")
```

```
  listings_temporal_final <- rbind(temporal_temp_listings_2)
  }
listings_final <- rbind(listings_spatial_final, listings_temporal_final)
listings_final <- listings_final %>% mutate(Source = "Listings")

#####################################
# Pivotieren der Wörter in Satzgruppen
#####################################
# Jede Beschreibung enthält eine eindeutige ID
# Nach dieser ID sollen die einzelnen Wörter zu einem Satz (ohne
Satzzeichen)
# zusammengefügt werden
# Der Code ist, bis auf die Variablennamen gleich wie weiter oben
# Am Ende wird der finale Dataframe als CSV exportiert und händisch
codiert
listings_pivot_id_1 <- as.character(listings_final$id_join)
# Teilen der Zeilen IDs -> dadurch ist später eine Zuordnung zu
zusammengehörigen Sätzen möglich
listings_pivot_id_1 <- strsplit(listings_pivot_id_1, split ='.',
fixed=TRUE)
# Maximale Länge/Breite der Liste bestimmen
len <- max(lengths(listings_pivot_id_1))
# Wenn Liste kürzer als maximale Länge, dann mit NAs füllen
listings_pivot_id_1_cols <- lapply(listings_pivot_id_1, function(l) c(l,
rep(NA, len - length(l))))
# Liste in einen Dataframe umwandeln
listings_pivot_id_temp <-
data.frame(matrix(unlist(listings_pivot_id_1_cols),
nrow=length(listings_pivot_id_1_cols), byrow=T))

listings_final_pivot_temp <- cbind(listings_final,
listings_pivot_id_temp)

listings_final_pivot <- listings_final_pivot_temp %>%
  group_by(id, type_of_recognition, Source, X1) %>%
  summarise(sätze = paste(word, collapse=" "))

# Export der relevanten Bewertungen

FINAL_reviews_listings <- rbind(reviews_final_pivot,
listings_final_pivot)
write.csv2(FINAL_reviews_listings, "reviews_listings_to_encode.csv",
row.names = F)
```

Anhang III Auflistung der detektierten Sätze mit potenziellen Raumbezug und dessen Ergebnis zur händischen Kodierung

Die Art der Angabe beschreibt, ob es sich um eine zeitliche Angabe (Z) oder eine räumliche Angabe (R) handelt. Als Quelle gilt entweder die Bewertung eines Gastes (Rating) oder die eigene Beschreibung des Gastgebers (Hosts).

ID	Art der Angabe	Quelle	ID vom Satz	Inhalt des Satzes	Code	Rau [m]	Zeit [min]	POI
1	Z	Rating	368	station so stephansplatz which is the center of the centrum is just 15 mins away i stayed there with my parents and as my mother	1			
1	Z	Rating	377	lage ist sehr bequem 20 minuten zu fuss bis zum zentrum und 30 min bis zum schloss schoenbrunn die wohnung selbst ist sehr nett da	2		30	Schloss Schönbrunn
1	Z	Rating	379	après une journée de visites la situation est exceptionnelle on est à 10min de stephansdom 10 min de schonbrunn et à 1 min de la de visites la situation est exceptionnelle on est à 10min de stephansdom 10 min de schonbrunn et à 1 min de la gare station de métro tramway et bus aeroport le	3			
1	Z	Rating	388	great location very comfortable and within walking distance of the city center 20 min shopping area 10 min and bakery grocery store restaurant public transportation and city bikes 5 min comfortable beds for all 5 of us will definitely recommend this	1			
1	Z	Rating	401	close to the underground you can reach city center in less than 10 minutes our stay in wien couldn't be better thank you peter and	1			
1	Z	Rating	409	have to be quite good friends it's in a fantastically convenient position 2 mins from westbahnhof which is a transport hub you can be in the city centre in 10 mins i would 100 recommend it to anyone wanting a good value	2		2	Westbhf.
1	Z	Rating	412	least 1 additional person the apartment is also extremely well located literally 2 min walking to the train station where you can get the bus	1			
1	Z	Rating	428	this apartment is good location then house is big enough for 3 4person and warm clean just like home i would stay again here	0			
1	Z	Rating	429	die	0			
1	Z	Rating	437	flat's location is very close to central station it takes just 5 10min so there are supermarket	0			

				and cafe bakery as well in the central			
1	Z	Rating	443	explications très claires et utiles le bus arrive devant la gare à 5 minutes à pied du logement peter était là pour nous accueillir il	3		
1	R	Rating	444	all four of us that peter said that parking is free after 8pm but as it turns out it is only free after 10pm to after 8pm but as it turns out it is only free after 10pm to be honest i hoped that he will pay the fine because us false information but he didn't he remembers that he told us 10pm from the start aside from this we had a great stay	0		
1	R	Rating	445	our	0		
1	Z	Rating	446	peter when we arrived and he came downstairs to greet us in 2 mins he was also helpful in providing info on the city overall	0		
1	Z	Rating	451	more adventurous you can also walk to the central area in about 25 minutes special thanks to peter for turning up in the rain to	1		
1	R	Rating	456	apartment	0		
1	R	Rating	457	peter was super kind to wait for our late arrival after 9pm to check us in and even had some milk in the fridge	0		
1	Z	Rating	458	was very clean and everything was provided it is also well located 3 minutes away from metro tram and a major train station there are	2	3	Westbhf.
1	Z	Rating	462	after my call peter arrived just 2 minutes after and hosted us to the apartment he was really nice	0		
1	Z	Rating	466	bus station which was direct to westbahnhof the apartment was less than 5 mins away on our departure there was also an airport bus from	1		
1	Z	Rating	479	could walk to the very center of vienna if you'd like about 30min strongly recommend to anyone who would like to explore this great city	1		
1	Z	Rating	488	our queries or messages was superfast the flat is located just a 2 min walk from westbahnhof and is exactly as the pictures describe it	2	2	Westbhf.
1	Z	Rating	493	it is very close to the train station we spend less than 10 minutes walking from the train station to the house then the house	0		
1	Z	Rating	500	to a main train station there are some good restaurants within 15 20 mins walk and a good bakery five mins away so it makes a good base to explore a great	1		
1	R	Rating	502	quiet building	0		
1	R	Rating	503	everything went well peter waited for us we arrived at 9pm and he was at the apartment when we	0		

				arrived he explained everything			
1	Z	Rating	506	the flat is in a great location 2 minutes to trams and train station very warm and spacious beds were	1		
1	Z	Rating	533	but the apartment was perfect very close to downtown probably about a 30 min walk we went running by the castle it was gorgeous the	1		
1	Z	Rating	543	sehr wohl gefühlt die lage ist toll ausgehviertel sind in gehdistanz max 20min zu fuss und per tram u bahn ist fast alles in weniger fuss und per tram u bahn ist fast alles in weniger als 15min zu erreichen	1		
1	Z	Rating	544	beautiful apartment that is well positioned in the city of	0		
1	Z	Rating	565	we could access with the subway it was a short subway ride 10 mins to city center peter gave easy directions that made getting around	0		
1	Z	Rating	579	right next to the main metro train website hidden by airbnb takes 1 min to get there peter is an amazing host he gave us	1		
1	R	Rating	616	sur les bons plans de vienne un appartement génial à même pas 100m de la station de métro une rue commerçante à 150m un super même pas 100m de la station de métro une rue commerçante à 150m un super appartement pouvant accueillir de nombreuses personnes tres content d'être tombé	3		
1	Z	Rating	641	walk to the train station that takes you to downtown in 7 10 minutes peter communicates quickly and is super helpful we highly recommend	0		
1	Z	Rating	648	close to public transit allowing easy access to the city center 15 20 mins via tram subway peter was an excellent host and his suggestions	0		
1	Z	Rating	650	in vienna well located near westbanhof close to many public transports some 20 minutes to city center well equipped spacious and bright the only downside	0		
1	Z	Rating	652	everywhere in city centre also there is metro station which is just 3 min walk away from home	1		
1	Z	Rating	657	where you have access to metro trams trains and airport bus just 5 minutes walk to mariahilf strasse where you can find tons of cafes	1		
1	Z	Rating	689	excellent very easy assess to every possible location public transport is just 4 min away even if you are in the mood for a walk	1		
1	Z	Rating	694	check in and check out is super easy location is great just 5 mins from westbahnof station wide	2	4	Westbhf.

				space we are four ppl and don't			
1	R	Rating	697	to metro	0		
1	R	Rating	698	la llegada al departamento es muy sencilla se encuentra a 150m de la estación westbanhof una de las principales de vienna el edificio	3		
1	Z	Rating	706	hours before take off the private transfer is just convenient good location 5min walk to train station trams buses good view of sunrise 7 to walk to train station trams buses good view of sunrise 7 to 10 min wall to three supermarkets and a vegan supermarket clean bedsheets and	1		
1	Z	Rating	749	mention location is perfect we could get pretty much everywhere within 20 25 mins by underground peter welcomed us really kindly and he also showed	0		
1	Z	Rating	762	also stunning and is right next to travelers shack vienna only a 7 min metro ride to the heart or vienna	0		
1	Z	Rating	770	apartment was huge 2 min walk to metro station clean peter was there to receive and	0		
1	Z	Rating	772	you don’t hear it this gives you a quiet stay that is 15 minutes or less from almost all parts of the city on a	1		
1	R	Rating	774	zentral aber trotzdem ruhig	0		
1	R	Rating	775	<U+4E00><U+5207><U+90FD> <U+8D85><U+7EA7> <U+68D2> alles super	0		
1	R	Rating	776	goede plek op 200m van westbahnhof 5 min lopen groot wnkelgebied en ov op 2 min	3		
1	Z	Rating	776	goede plek op 200m van westbahnhof 5 min lopen groot wnkelgebied en ov op 2 min lopen zeer schoon appartement en goede bruikbare locale tips super	3		
1	Z	Rating	790	great spacious apartment excellent location just 5 7 minutes walking distance from the westbahnhof railway station which is also a	2	5	Westbhf.
1	Z	Rating	794	<U+C788><U+C744> <U+AC83><U+B4E4><U+C740> <U+B2E4> <U+C788><U+C5B4><U+C11C> <U+D3B8><U+D558><U+AC8C> <U+BCF4><U+B0C8><U+C2B5><U+B2C8><U+B2E4>	0		
1	Z	Rating	795	an excellent apartment with all amenities 5min from the city center gym supermarket bus station train restaurant shops the	1		
2	Z	Host	3	the airport there is a direct train connection a ride takes about 20 minutes and costs ca 4 euros	1		
2	Z	Rating	367	for comfort living but the most advantage of it is it's place 7 min	2	7	Oberes Belvedere

				walk from belvedere so if u want to spend time in			
2	Z	Rating	369	us easy to follow instructions to the apartment which is very close 2min walk to transport the location was wonderful for walking to sights and transport the location was wonderful for walking to sights and an easy 10min walk into the city center the apartment was one of the best	1		
2	Z	Rating	380	surprisingly well equipped the tramway brings you to karlsplatz in less than 10 minutes the train is one stop away from landstrasse wien mittte i	0		
2	Z	Rating	390	direct to the airport and to mitte the latter in less than 5 mins and there are also convenient trams eg no 71 takes you	0		
2	Z	Rating	393	we liked the appartment because it was in a real viennese neighborhood 15 minutes from downtown by tram and subway salvatore has a great system	1		
2	Z	Rating	395	a sweet little space and was perfect for 3 people it's a 15 min walk to some must see sights and is a 2 min walk to a train station so getting around is really easy	1		
2	Z	Rating	407	great the location was good walking distance to the center and ca 30 min to the naschmarkt which is a visit worth we got very	2	30	Naschmarkt
2	R	Rating	408	could go the airport in time because the earliest train was 4 33am the building is a little bit old but the room has been	0		
2	Z	Rating	428	is good near to the city centre and the station is just 5 minutes away from the flat salvatore is a great host helpful and	1		
2	Z	Rating	464	you need in the apartment the location is not bad 20 to 30 minutes walk to the national opera house but there's a tram nearby	2	25	Wiener Staatsoper
2	Z	Rating	466	located if you enjoy walking every interests sites is within a max 20 minutes walk grocery store billa and casual places to eat or drink	1		
2	Z	Rating	498	to the city centre but we choose to walk which took about 20 mins the room was clean and well equipped we would definitely recommend	1		
2	Z	Rating	501	perfect size for two people the location is great and only a 15 min walk to the center of vienna it's very close to the	1		
2	R	Rating	503	a reasonable 20min walk to just about anywhere in the 2nd or 1st district albeit a cold walk basic needs were met with this apartment	1		

2	Z	Rating	503	a lot of walking from the apartment and it was a reasonable 20min walk to just about anywhere in the 2nd or 1st district albeit warm during the cold few days we stayed in vienna about a 5 min walk from the metro station with a number of grocery stores	1		
2	Z	Rating	505	from the airport the apartment is well located closed to belvedere and 15 mins walking from the underground station house was nice well equipped and	1		
2	Z	Rating	515	is beautiful with bars restaurants and belvedere park there is one supermarket 2 min walking metro station is 3 min walking the bed is very comfortable and big i would recommend	1		
2	R	Rating	519	neighbourhood with more trees and a bit more multicultural life than the 1st which is very close 10 minutes walk you can also take the	1		
2	Z	Rating	519	it is super easy and fast to reach from the airport around 30 minutes by train than 5 min walk max to the appartment the appartment is located in the a bit more multicultural life than the 1st which is very close 10 minutes walk you can also take the tram 2 3 min walk and there is a city bike station around the corner	1		
2	Z	Rating	535	staying here the train station to go to the airport is a 5 min walk	1		
2	Z	Rating	547	ausgestattet was man für einen kurzen aufenthalt benötigt und vom hbf in 10 min zu erreichen	2	10	Hauptbahnhof
2	Z	Rating	561	l'appartement est bien situé proche de la gare de rennweg et à 20min à pied du centre le principe pour avoir la clé est magique	3		
2	Z	Rating	585	lovely flat conveniently located next to the belvedere area 5 minutes walking distance from main transportation and 15 20 minutes away from city center on foot easy access from the airport	1		
2	Z	Rating	595	us the apartment is clean quiet and warm it is only a few minutes from the belvedere museum which is definitely worth a visit it is about 15 20 minutes walk from the old town which may be a bit far	1		
2	Z	Rating	601	lovely apartment in good location approx 3 mins from rennweg s bahn station where there is a lidl burger quick access to the city city centre is also accessible on foot 5 min walk through botanical gardens to belvedere	2	3	Rennweg
2	Z	Rating	627	walk to the wiener staatsoper from the apartment within about 20minutes or 15min by bus also there's a big park beside it is also	2	20	Wiener Staatsoper

				close			
2	Z	Rating	635	of a tram that takes you to the centre of vienna in 10 min there are big supermarkets nearby to buy groceries the self check	1		
2	Z	Rating	644	peut paraître quelque peu excentré au premier abord mais on est à 30min à pied du centre et 10min en tram je conseille vraiment ce premier abord mais on est à 30min à pied du centre et 10min en tram je conseille vraiment ce logement	3		
2	Z	Rating	645	studio parfait pour visiter vienne	3		
2	Z	Rating	653	anreist zur innenstadt zu fuß stephansdom ist es allerdings ein stück ca 30 min zum schloss belvedere ca 10 min in direkter nähe befindet sich die station rennweg sowie ein city	2	10	Oberes Belvedere
2	Z	Rating	653	anreist zur innenstadt zu fuß stephansdom ist es allerdings ein stück ca 30 min zum schloss belvedere ca 10 min in direkter nähe befindet sich die station rennweg sowie ein city	2	30	Stephansdom
2	Z	Rating	657	need and the shower is amazing very powerful the location is perfect 5 mins walk from the belvedere museum and botanical garden and very close and restaurants it's very easy to get a tram from a stop 2 mins walk from the apartment despite being right in the centre the	2	5	Oberes Belvedere
2	Z	Rating	667	follow and very useful about the local area it is a 2 3min walk to the botanical gardens which you can walk through to the	2	3	Botanischer Garten
2	Z	Rating	668	the main attractions in vienna we found that it was within approximately 20 minutes walking distance to almost all sights communication with ramona was excellent	1		
2	Z	Rating	676	single traveller quick and easy check in procedure with clear instructions about 20 mins walk into the heart of vienna thanks for having us	1		
2	R	Rating	681	and kindly arranged for us to check in a little earlier than 2pm would recommend to any individual or couple staying in vienna	0		
2	R	Rating	682	everything ok	0		
2	Z	Rating	688	about it lots of local shops for amazing breads and less than 10 min walk to the botanical gardens and the belvedere art gallery fantastic	1		
2	Z	Rating	690	site was more than a 35 minute walk away with several within 20 minutes the neighborhood had all the main conveniences grocery stores banks etc	1		

3	Z	Host	6	public transport is unbeatable from this u1 station nestroyplatz is a 2 minute walk away u4 schwedenplatz is 8 minutes away and u2 praterstern is 9 minutes away and the main tramway line 1 that goes around the	2	2	Nestroyplatz
3	Z	Host	6	public transport is unbeatable from this u1 station nestroyplatz is a 2 minute walk away u4 schwedenplatz is 8 minutes away and u2 praterstern is 9 minutes away and the main tramway line 1 that goes around the	2	9	Praterstern
3	Z	Host	6	public transport is unbeatable from this u1 station nestroyplatz is a 2 minute walk away u4 schwedenplatz is 8 minutes away and u2 praterstern is 9 minutes away and the main tramway line 1 that goes around the	2	8	Schwedenplatz
3	Z	Rating	1431	donnent sur une cour intérieure la localisation de l'appartement est idéal à 2 minutes de la bouche de métro nestroyplatz et le quartier est un	3		
3	Z	Rating	1439	places to go the flat is very near the underground less than 5 min walking we really recommend staying at this place	1		
3	Z	Rating	1445	a crib and allow our child to roam around while located only 15 minutes away from stefansdom by foot the area is very quiet and	2	15	Stephansdom
3	Z	Rating	1462	apartment is located is quiet and the u bahn 1 is just 5 minutes walking unfortunately i couldn't meet cristina but she was always available	1		
3	Z	Rating	1470	and shop the apartment is very close to the river and thus 10 minutes on foot by the city center with all possible public transport	1		
3	Z	Rating	1501	station shops and restaurants the walk to the city center is about 20 min long and really nice overall we had a great experience in	1		
3	Z	Rating	1504	with rain it's a simple apartment that has been very well furnished 10 min walk to wien praterstern station about a 25 minute walk to with a couple of good bars and restaurants close by within a 5 minutes walk there's also a great market around the corner if you get there early for breakfast would definitely recommend	2	10	Praterstern
3	Z	Rating	1512	spent two nights i cristina's flat the place is nicely located within 15 min from the main attractions and old city center thw flat had	1		
3	Z	Rating	1601	tout est accessible a pied sinon la station de metro est a 5min propre l'hôte est ponctuelle sans soucis de communication ceci dit l'appartement est quartier tranquille a 10 min du centre ville	3		

				tout est accessible a pied sinon la station équipements pas de soucis tout y était voilà	
3	R	Rating	1603	very good restaurant just across the road and a small supermarket within 100 meters we walked to lots of venues for the fun of finding	1
3	Z	Rating	1603	all mentioned in guide books you can be in the city in 10 minutes	1
3	Z	Rating	1606	a good located flat only 15 minutes walk to the city center with all the main sights only	1
3	Z	Rating	1609	good location about 15 minutes walk to the centre and the apartment is in a nice	1
3	R	Rating	1614	<U+C804><U+CCA0><U+C5ED><U+C774><U+B791> <U+D2B8><U+B7A8><U+C815><U+AC70><U+C7A5><U+C740> 3 5 <U+BD84> <U+AC70><U+B9AC><U+C5D0> <U+C788><U+C5C8><U+C2B5><U+B2C38><U+B2E4>	0
3	R	Rating	1615	<U+5730><U+70B9> <U+79BB> <U+666F><U+70B9> <U+7EA6> 2 3km <U+9700><U+8981> <U+4E58><U+5750> <U+5730><U+94C1> <U+4F46> <U+623F><U+95F4> <U+5E72><U+51C0> <U+65B9><U+4FBF> <U+8BBE><U+65BD> <U+9F50><U+5168> <U+623F><U+4E1C> <U+975E><U+5E38> <U+70ED><U+5FC3>	0
3	Z	Rating	1630	un vélo super pratique pour rejoindre le centre accessible à pied aussi 15min des restos sympas dans le coin un joli marché bref pour une	3
3	R	Rating	1655	our stay in this cozy apartment however be prepared to walk approximately 1 mile or take public transit to most attractions	1
3	Z	Rating	1681	apartment was clean had all the necessities and was a nice 20 30min walk into the city center definitely recommend	1
3	Z	Rating	1682	cristina’s place is an amazing	0
3	Z	Rating	1693	clean and with a lot of amenities great location everything is within 10min walk would 100 recommend	1
3	Z	Rating	1694	<U+C695><U+C870><U+AC00> <U+C788><U+C5B4><U+C11C> <U+C88B><U+C558><U+C5B4><U+C694> <U+BB54><U+AC00>	0

				<U+D070> <U+C6D0><U+B8F8> <U+B290><U+B08C><U+C785> <U+B2C8><U+B2E4> <U+B113><U+C5B4><U+C6A9>				
3	Z	Rating	1695	easily the best located air bnb i have used 5 minutes walk from nestroyplatz u1 taborstrasse u2 and 10 minutes from schwedenplatz u4 could walk all the way to the museums quarter in 30 minutes and stephansplatz in 15 a simple comfortable base for exploring the	2		5	Nestroyplatz
3	Z	Rating	1695	easily the best located air bnb i have used 5 minutes walk from nestroyplatz u1 taborstrasse u2 and 10 minutes from schwedenplatz u4 could walk all the way to the museums quarter in 30 minutes and stephansplatz in 15 a simple comfortable base for exploring the	2		10	Schwedenplatz
3	Z	Rating	1706	great location in a quiet neighbourhood very close to public transport 5 min walk to the nearest tram station and in 2 or 3 back to vienna	1			
4	Z	Rating	368	the museumsquartier subway stop a 3 minute walk from the kunsthistorische museum 15 minutes from the opera when we walked into the place i said	2		3	Kunsthist. Museum
4	R	Rating	368	to the apartment which is really nice and incredibly well located about 10 m from the museumsquartier subway stop a 3 minute walk from the	2	20		MQ Wien
4	Z	Rating	368	the museumsquartier subway stop a 3 minute walk from the kunsthistorische museum 15 minutes from the opera when we walked into the place i said	2		15	Wiener Staatsoper
4	Z	Rating	381	super location direkt am mq quartier nur 10 min zum stephansdom übergabe war perfekt und die wohnung sehr gut eingerichtet	2		10	Stephansdom
4	Z	Rating	400	at the same place there is a supermarket it's less than a 45min trip from vienna airport to the apartment using public transport the apartment	0			
4	Z	Rating	414	minor crisis which erich helped resolve and was by our sides within 30 minutes what service the location is really perfect being so close to were on the street with the secession building and a short walk 20 minutes to st stephens square we took the english speaking taxi from	0			
4	Z	Rating	437	clean modern fresh towels were provided close to the u2 and about 15 min walk 25 if you got lost like i did to u1	1			

4	Z	Rating	445	ga en rask omvisning i leiligheten vi tok taxi fra flyplassen ca 25 min utenom rushtrafikk leiligheten har perfekt beliggenhet hele bykjernen tilgjengelig innen gåavstand	3		
4	Z	Rating	452	the images and very convenient close to lots of museums and just 30 minutes to the airport there is also a grocery downstairs which is	0		
4	Z	Rating	454	a great location right by the museums and only a short 10 15min walk to the centre of the city definitely a nice spot for	1		
4	Z	Rating	455	the apartment the location is really great it is just in 3 5 minutes from hofburg secession and many other museums and buildings just in	2	5	Secession
4	Z	Rating	480	a metro station and lots of museums also hofburg palace is like 5 minutes walk erich is very kind and helpful the flat is really	2	5	Hofburg
4	R	Rating	491	us there is so much noise from the construction beginning at 6 30am and lasting until 7pm this might have been the worst part and much noise from the construction beginning at 6 30am and lasting until 7pm this might have been the worst part and erich didn't seem aware	0		
4	R	Rating	500	<U+ADF8><U+B9AC><U+ACE0> <U+AC74><U+C870><U+AE4C><U+C9C0> <U+AC00><U+B2A5><U+D588><U+ACE0> <U+C138><U+D0C1> <U+C0AC><U+C6A9><U+BC95><U+AE4C><U+C9C0> <U+C54C><U+B824><U+C8FC><U+C5B4> <U+C88B><U+C558><U+C2B5><U+B2C8><U+B2E4> we arrived very late after 11pm but he waited for us he explained all the things and also	0		
4	Z	Rating	517	hôte n'est pas particulièrement accueillant et arrangeant on ne l'a vu que 5 min pour qu'il nous remette les clés	3		
4	R	Rating	518	the only downside was the very loud construction work which started around 6.30am making sleep impossible from this time onwards however this is clearly something	0		
4	R	Rating	531	location for those wishing to see the old town with everything at 10m walking metro and bus just around the corner make it even more	1		
4	Z	Rating	536	distributeur d'argent la station de métro et le quartier des musées à 2 minutes à pieds tout était parfait	3		

				dans cette location nous sommes certains			
4	Z	Rating	539	great apartment and amazing location right in the city center 5 min walk to albertina museum shops and cafes right around the corner	2	5	Albertina
4	Z	Rating	552	large clean and calm flat the ideal place to visit vienna 10 minutes walk to hofburg	2	10	Hofburg
4	Z	Rating	564	i've only met erich for 5 minutes but everything went very smoothly	3		
4	Z	Rating	568	est très très bien situé à 2 min du metro museum quartier 5min de karlplatz et volkstheater il était ponctuel et tout était simple d'accès bon séjour à vienne l'appartement d'erich est très très bien situé à 2 min du metro museum quartier 5min de karlplatz et volkstheater il était	3		
4	R	Rating	587	lot of sights are on the walking distance public garage is only 300m away 8 daily very cheap for vienna good value for your money	1		
4	Z	Rating	588	was accommodating when i told him i would need to check in 30 minutes earlier	0		
4	R	Rating	590	empfehlen alle wesentlichen sehenswürdigkeiten sehr gut zu fuss erreichbar und metro nur 100m entfernt	1		
4	R	Rating	591	erich was very responsive to texts and though our plans changed	0		
4	Z	Rating	606	for short or long stay also parking house is very close only 2 minutes from appartment	1		
4	Z	Rating	625	excellent location 3 mins walk to the metro station and 10 mins to museum quartier plenty of cafes and restaurants within walking distance a supermarket is located just a min away the apartment is spacious with separate toilet and bathroom found	1		
4	R	Rating	633	peu plus de lumières mais pour un premier étage aux plafond de 4m c'était quand même assez luxueux on y reviendra	3		
4	R	Rating	634	excellent comfortable and clean	0		
4	R	Rating	647	erich but he's kind enough to let us keep our luggage until 2.30pm being a clean freak i found the floor was quite dirty and	0		
5	Z	Rating	10	centre of town the walk to city centre took us about 20 25 minutes definitely doable also located a short distance to belvedere museum 10 mins by foot we felt very comfortable in this apartment belinda herself	2	10	Oberes Belvedere
5	R	Rating	11	than expected and she was kind enough to check us in at 10.30pm throughout our stay she was	0		

				extremely quick to respond to any questions				
5	Z	Rating	13	all the facilities one would need the public transport access is about ten minutes walk away and connected us to all spots in the city about 15 minutes ride from the station	1			
5	Z	Rating	26	great time the tram station is very close and it only takes 8 minutes before you are in the heart of vienna the apartment is	0			
5	Z	Rating	33	and cozy to the center only two stops of the subway or 20 minutes by foot near the house is excellent cafe with a traditional cuisine.very tasty food and enormous portions for reasonable prices gasthaus is only 3 minutes by walk on the right side of the house near the it by blue and white colored tram.trip to baden will take about 50 min and cost only 12 euro for round trip andreas and belinda	0			
5	Z	Rating	34	for two people tram stop is really close and you are only 8 minutes away from the opera by tram 1 you can travel all	0			
5	Z	Rating	40	nearby which get you to the heart of the city in about 10 minutes appreciated the air conditioning very much since it was 37 c	1			
5	Z	Rating	42	apartment is either a 25 minute walk to the opera or a 10 min tram ride which was very easy we had a lovely time	2		25	Wiener Staatsoper
5	R	Rating	43	similar route and we got off at laurenzgasse station this station is 100 meters from belinda's appartment from laurenzgasse it is 3 4 stops to	2	100		Laurenzgasse
5	Z	Rating	43	kettle a toaster and a microwave a huge grocery store eurospar is 5 minutes walk on kliebergasse turn left at the end of jahngasse a convenient store a bakery and three very good restaurants are 3 minutes walk on the main road just on your way to the tram station laurenzgasse everything is around you and vienna's ring is 10 minutes away on the tram we never met belinda but her husband	1			
5	Z	Rating	47	kitchen coffee machine etc etc the location is great since its just 5 min walk from the metro line 1 and 62 which take you	1			
5	Z	Rating	49	neighborohood it is well located just few steps from transport which takes 10 minutes to get karlsplatz belinda organized the check in in all the	1			
5	Z	Rating	52	that will take you to the tourist area in just a few short minutes the ride to the main train station	0			

				for regional trains was perhaps 15 minutes the ride to the special train that takes you to the airport was maybe 10 minutes away though that train takes 30 minutes we chose to do that rather than take a taxi there			
5	Z	Rating	53	stop is less than a 10 minute walk and it takes about 10 minutes to get to the downtown area which is nice because you do enough walking as it is save your feet for a few minutes grocery shop on the way home from the tram stop it	1		
5	Z	Rating	66	in the photo description here clean spacious has ac public transportation is 4 minutes walk tram 62 takes you to all attractions in the city	1		
5	Z	Rating	69	arrival time due to traffic the apartment is in great location about 10 minutes from the subway station	1		
5	Z	Rating	74	the neighbour district favoriten beyond the railway it's free safe and just 10 min by foot	1		
5	Z	Rating	76	like on the photos the location is very good it takes just 5 minutes to get a tram so in 15 minutes you are in the vienna's center and it's close to hauptbahnhof	1		
5	Z	Rating	77	around the corner which takes you into the city centre in 10 15min we also walked into the city centre in 30 40min then main in 10 15min we also walked into the city centre in 30 40min then main train station is about a 20min walk the air con city centre in 30 40min then main train station is about a 20min walk the air con was great as we arrived in vienna in a few restaurants in the area and a supermarket is a 5 10min walk thanks belinda for letting us stay in your apartment	1		
5	Z	Rating	78	the flat	0		
5	Z	Rating	83	<U+304C> <U+3044><U+3044> <U+3067><U+3059> <U+306D> the apartment is very comfort it takes 20 30 minutes to walk to opera house you should use city trams	2	27,5	Wiener Staatsoper
5	Z	Rating	86	is really small and washing machine is outside the location is convinient 5 min from the tram stop 1 10 min to the city center and grocery shop is around belinda is	1		
5	Z	Rating	87	was very clean and cozy the city centre is reachable in 10 15min by tram belinda was very flexible with the check out hour and	0		
5	R	Rating	88	apartamento agradável e conforme descrito bem localizado a apenas 350 m da paragem do tram a belinda foi atenciosa e deu	3		

				nos			
5	Z	Rating	98	the location is good 20 minutes walking from central railway station nearby from the metro and tram stops so we could reach the city center in 15 minutes the neighborhood is quite with a few cafes and markets around	2	20	Hauptbahnhof
5	R	Rating	109	in vienna	0		
5	R	Rating	110	location is convenient to public transit trams and bus about 1km to interesting margareten neighborhood apartment is comfortable for 2 3 people it	0		
5	Z	Rating	127	wir haben der gastgeberin nescheid gegeben und die putzfrau ist innerhalb von 10 min gekommen und hat nochmal geputzt leider hat sie die gerüche nicht	0		
5	Z	Rating	129	recommended a relatively inexpensive parking garage however near kaiserplat which was located 20 mins away by car belinda and her husband were very helpful and	0		
5	Z	Rating	132	to anything you've probably experienced and trains trams tend to run every 5 minutes	0		
5	Z	Rating	137	der gegend gibt es einige nette restaurants und cafés alles ca 10 15 min zu fuß auch ein supermarkt ist direkt um die ecke und	0		
5	Z	Rating	142	located very centrally the main points if attraction can be walked in 40 minutes the place is small but exceptionally clean and cosy everything is	0		
5	Z	Rating	147	walkable to haubtbahnhof about 15 mins and very easy to find the building itself was quiet and	2	15	Hauptbahnhof
5	Z	Rating	149	bit far from the center in a residential area may take 20 30 mins to the sights but it is very quite and nice area	1		
5	Z	Rating	150	apartment is either a 30 minute walk to the opera or a 10 min tram ride which was very easy i definitely recommend you to	2	30	Wiener Staatsoper
5	Z	Rating	155	location we oftern walk to most of famous spots it takes about 40 minutes by foot to vienna golden hall and sr stephen's cathedral but	2	40	Musikverein
5	Z	Rating	165	tram station which take you to the city centre in just about 8 minutes there is a grocery store and a few places where you	0		
5	Z	Rating	175	appartement à 45 minutes du centre ville très bien desservis et qui permet de découvrir	3		
5	Z	Rating	185	des transports en communs bien que les trajets sont faisables à pied 20 min du centre	3		
5	Z	Rating	192	very nice apartment good position 15 minutes from city center	1		

5	Z	Rating	194	bien situe proche des transports pour découvrir vienne bus tram 7 a 10mn	3		
5	Z	Rating	195	super preis leistungs verhältnis es war alles vorhanden und hat funktioniert mit	0		
5	Z	Rating	198	vienna we often walked into the city centre which is about 25 30 mins or easily caught the tram which is very close to the	1		
5	R	Rating	199	burner electric stove top the location was great too only about 1 2 miles from all the attractions in vienna we walked downtown and to	0		
5	Z	Rating	200	great place to stay in vienna location is great just 2 minutes away from the tram station totally worth it for the price	1		
5	Z	Rating	207	within walking distance of city center about 30 minutes central station is only a 15 20 minute walk there is	2	17,5	Hauptbahnhof
5	Z	Rating	208	all the location is great the apartment is a couple of blocks 5min walk to the trams which take you directly to the opera house house and ringstraße the apartment is also only a few short blocks 10min walk to the main streets in the neighborhood where many great restaurants	1		
6	Z	Host	2	our panda twins u3 you can reach with tram 10 or in 10 minutes foot walk with this subway you have a fast and easy you want to go in vienna trams and subways go every 3 5 minutes during the day in the evening you have to wait a	1		
6	Z	Rating	212	meinem zimmer die gastgeberin sehr sympathisch und hilfsbereit die straßenbahn ist in 2 min zu erreichen geschäfte in der nähe für meinen nächsten wien aufenthalt	0		
6	Z	Rating	216	straße auch ziemlich ruhig entfernung zur innenstadt stephansplatz absolut im rahmen ca 7 min fußweg zur u bahn dann 13 minuten u bahn kann ich	1		
6	Z	Rating	217	the time contactable near to the tram and to the subway station 30 min to the center aprox the room is comfortable quiet and very	0		
6	Z	Rating	223	really recommend ulrike's apartment it's not downtown area but you are about 10 minutes walking from ottakring u bahn station it's a huge apartment and	2	10	Ottakring
6	Z	Rating	232	connectivity to the town by bus and tube it took us roughly 30 min to reach the city centre there was ample place below the	0		
6	Z	Rating	238	lovely and clean place in a convenient location this place is about 25 minutes from the central	0		

				of vienna by tram number 44 30 minutes away from westbahnhof vienna train station and 1 hour away from			
6	Z	Rating	240	sharing an apartment i felt i had enough privacy it is about 20 minutes to the center of vienna transportation is easily accessible thank you	0		
6	R	Rating	257	very kindly sincerely i was scheduled to arrive at vienna's airport around 10pm i traveled alone so i was a bit worried about getting to	0		
6	Z	Rating	257	exactly as the pictures you see at the website close to tram 4 minutes walk also you can walk to metro a bit further the	1		
6	Z	Rating	258	centre of the city is close handy and it only takes about 15 mins plenty of places to eat close by also could not recommend	1		
6	R	Rating	275	does not matter at all the tram to the center is at 100 meters the closest metro station ottakring u 3 at 10 minutes walk many supermarkets within 200 meters where there are also possibilities to have breakfast lunch or dinner	2	10	Ottakring
6	Z	Rating	275	is at 100 meters the closest metro station ottakring u 3 at 10 minutes walk many supermarkets within 200 meters where there are also possibilities	2	10	Ottakring
6	Z	Rating	277	ourselves like at home.the flat is very close to the downtown appr 20 minutes by tram and tube so we took long walks all day	0		
6	Z	Rating	279	has a big desk and it's very quiet the neighborhood is about 20 minutes by tram from the city center but it's a chance to	0		
6	Z	Rating	283	für mich war der bequemste zu fuß zur tramhaltestelle nr 44 ca 5 min dann zum schottentor 16 min dann ist man schon im zentrum steffl rathaus usw gut zu	1		
6	Z	Rating	285	in ottakring which takes you straight into the inner city in about 15 minutes so you are well connected too all in all i enjoyed	0		
6	R	Rating	290	zu ihr nach wien fahren	0		
6	R	Rating	291	ulike was lovely and welcoming even at 9pm after she finished work and we were a last minute booking	0		
6	R	Rating	292	abbiamo	0		
6	Z	Rating	298	is a bit further away from the city center took me about 40 minutes nevertheless it is a great value and i would definitely stay	0		
6	Z	Rating	311	44 che passa a 200 mt ogni pochi minuti impiega meno di 15 min per arrivare a shottentor che è già centro e da cui	3		
6	Z	Rating	322	days there is a train station and a	1		

				tram stop at a 5 minutes walk form the flat from where you can easily reach any way to approach the university was by train it took more than 40 minutes and when i explained ulrike she recommended me to take the tram 20 minutes by the way make sure that during your time there you			
6	Z	Rating	323	to the tiredness from all our travels d the apartment was about 15 minutes from what i remember outside the city but the public transport	1		
6	Z	Rating	326	the city centre trams 10 and 44 can take you there within 20 minutes what matters is that her lovely house is in a very	0		
6	Z	Rating	346	she prepared some for me the flat was wonderful and clean around 15 min to reach the city center by the metro u3 everything was	0		
6	Z	Rating	354	person her home is quiet pretty and comfortable it is located about 25 minutes from the center of town with easy access via multiple public and getting to know this neighborhood great austrian indian turkish and greek restaurants minutes away by foot grocery stores one block away and then across	0		
6	Z	Rating	365	so habe ich z.b dieses mal die strassenbahnlinie 2 fussweg dorthin max 5min ausprobiert die direkt in die stadt fährt	1		
6	Z	Rating	366	ulrike gave me a very	0		
7	R	Host	5	bester zugang zu den öffentlichen verkehrmittel da die wohnung im zentrum gegenüber	0		
7	Z	Rating	855	did not see her her place is in the center of vienna 3 min walk to st stephen's cathedral and the metro her place is	2	3	Stephansdom
7	R	Rating	877	eating places are just stone's throw from the property metro is just 200 mts purchase a day ticket under 7 euro n travel anywhere in	1		
7	Z	Rating	909	it was clean and tidy which is always nice it is only 5min from the metro and 12min walk from the airport train link station which is always nice it is only 5min from the metro and 12min walk from the airport train link station	2	12	Landstraße-Wien
7	Z	Rating	910	elisa’s apartment was very comfortable	0		
7	Z	Rating	928	elisa's place is at the center of everything just a 5 15 mins walk to all the sights and shopping street the apartment is	1		
7	R	Rating	949	this place is a fabulous location it's about 200 feet from the center of vienna and almost all the great sights	1		
7	Z	Rating	957	is a very kind and helpful host the	2	2	Stephansdom

				flat is really central 2 minutes walk from the cathedral but in a quiet area beautiful and				
7	Z	Rating	959	which could be critical for those with mobility issues it is around 5 minutes on foot to reach the closest u bahn metro station and within these 5 minutes there are also supermarkets pharmacies and various restaurants and boutiques this	1			
7	R	Rating	979	messaged to make sure we were ok this was a nice touch	0			
7	R	Rating	980	1st ever visit to beautiful vienna we fell in love with city and	0			
7	Z	Rating	987	apartment was perfectly clean and it was such a perfect location a 3 min walk to the metro and the cathedral	2		3	Stephansdom
8	R	Rating	1481	and bright high ceiling centrally located near the pilgramgasse u bahn station 40 meters in a street that has all ammenities shops cafés banks and	2	40		Pilgramgasse
8	Z	Rating	1485	directly across the street public transportation 1 minute walk from the apartment 5 minutes train ride to city center check in and out extremely simple	1			
8	Z	Rating	1500	15 20 mins walk to the city center right by the metro the room	1			
8	R	Rating	1515	<U+623F><U+5B50>	0			
8	R	Rating	1516	nous avons beaucoup apprécié la situation parfaite de cet appartement à 50m d'un arrêt de métro sur une ligne à 5 min de karlsplatz	3			
8	Z	Rating	1516	arrêt de métro sur une ligne à 5 min de karlsplatz et 15min de schwedenplatz d'où on peut aller n'importe où à vienne la présence cet appartement à 50m d'un arrêt de métro sur une ligne à 5 min de karlsplatz et 15min de schwedenplatz d'où on peut aller n'importe	3			
8	Z	Rating	1528	apartment with his accurate instructions when checking in location was also 10 15 minutes away from major tourist attractions with a car billa supermarket is	1			
8	R	Rating	1535	väsen om man har öppet fönster rekommenderas	3			
8	R	Rating	1536	the apartment is less than 100m from the underground station into the city centre it’s very clean and	2	80		Pilgramgasse
8	Z	Rating	1553	apartment is in a terrific location with pilgramgasse station no more than 5 minutes away and a taxi rank just around the corner we walked	1			
8	R	Rating	1580	elias apartment is big and comfortable and in a perfect location basically 50m from the	1			

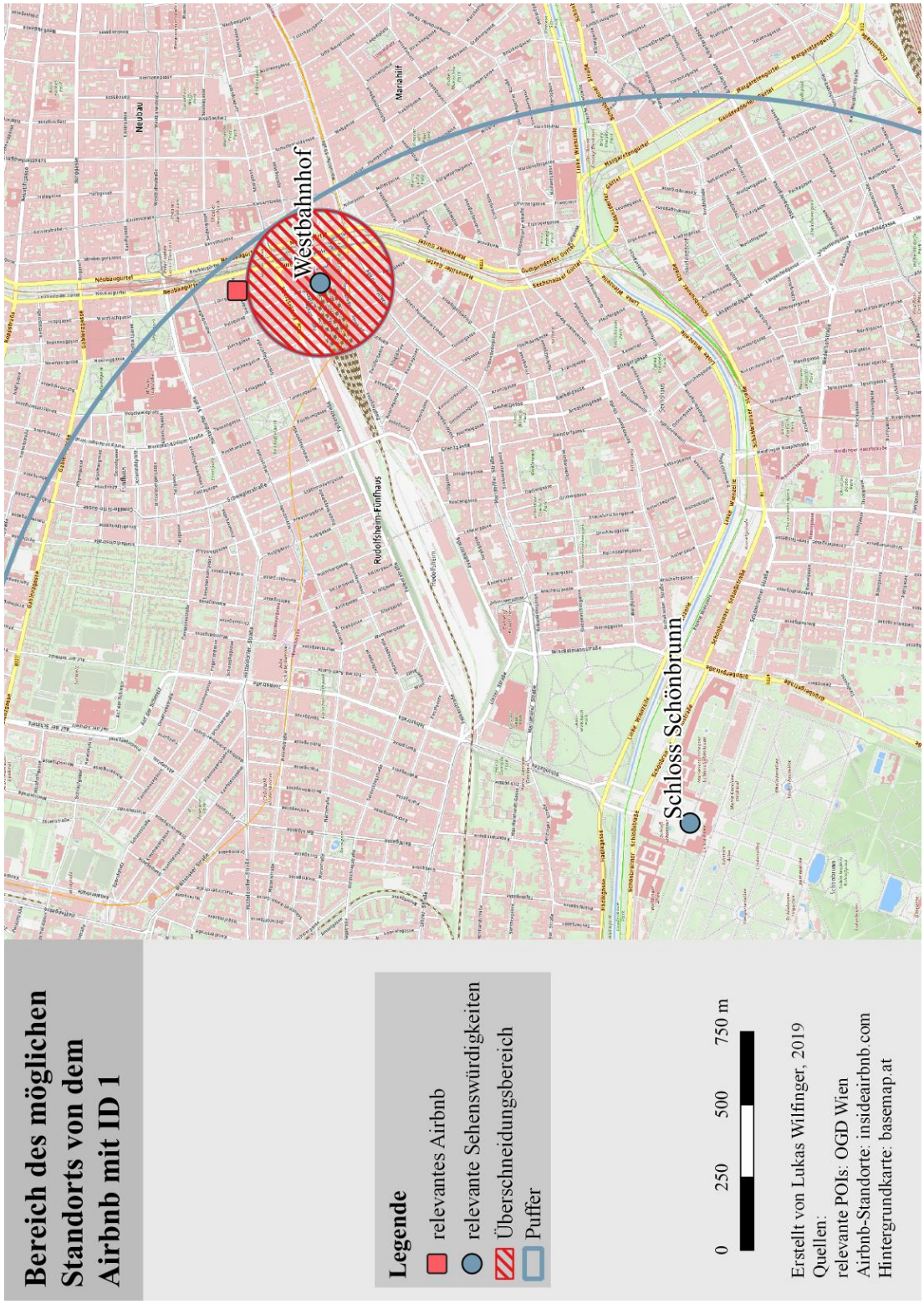
				bus and train stations the kitchen had everything we needed machines were in was very dirty but there was an accessible laundromat 200m or so from the place			
8	R	Rating	1581	great place in a cool spot	0		
8	R	Rating	1582	sehr	0		
9	Z	Host	6	metro u1 trams and busses in a 5 minutes walking distance main train station wien hauptbahnhof 5 minutes walking	2	5	Hauptbahnhof
9	Z	Rating	995	nightmarket museum quarter etc also very well communicated with the main station 5 min walking from where you can easy go to the airport also	2	5	Hauptbahnhof
9	Z	Rating	999	clean and beautiful it is close to vienna central train station for 10 minutes on foot it is also close to the underground station next	2	10	Hauptbahnhof
9	Z	Rating	1001	behind the house and on the other hand vienna's hauptbahnhof within a 5 minutes walk i will definitely come back thanks for everything	2	5	Hauptbahnhof
9	Z	Rating	1003	a lovely place really close to the vienna central station wienhauptbahnhof a 7 min walk where all the services can be public transportations including the restaurants grocerystore also close to the vienna city center by walk 15 20 min or with the public transportation at the vienna central station 10 min karin really takes the time to be sure we had all	2	7	Hauptbahnhof
9	Z	Rating	1009	city fast is really easy you could also walk there in 15 20 minutes what is there not to love i would strongly recommend this	1		
9	Z	Rating	1036	is an amazing host nice friendly and replies really fast place is 5 mins walk to wien station really quiet and nice room was big	2	5	Hauptbahnhof
9	Z	Rating	1041	to reach the apartment from the airport i spent just less than 30 minutes karin is the best host she’s very friendly and gave me	0		
9	Z	Rating	1054	very well maintained clean and it will feel you like home it's 10 mins walk from the metro so it's a convinent place in vienna	1		
10	R	Rating	1090	is erg comfortabel en compleet ingericht ligt op loopafstand van hoofdstation en 200 meter van metrostation keplerplatz je bent in een kwartier in de binnenstad	3		
10	Z	Rating	1103	2 3 minutes walk to the metro close to the hauptbahnhof for further transferring	1		
10	R	Rating	1104	the second floor 1 mun walk to a supermarket which opens from	0		

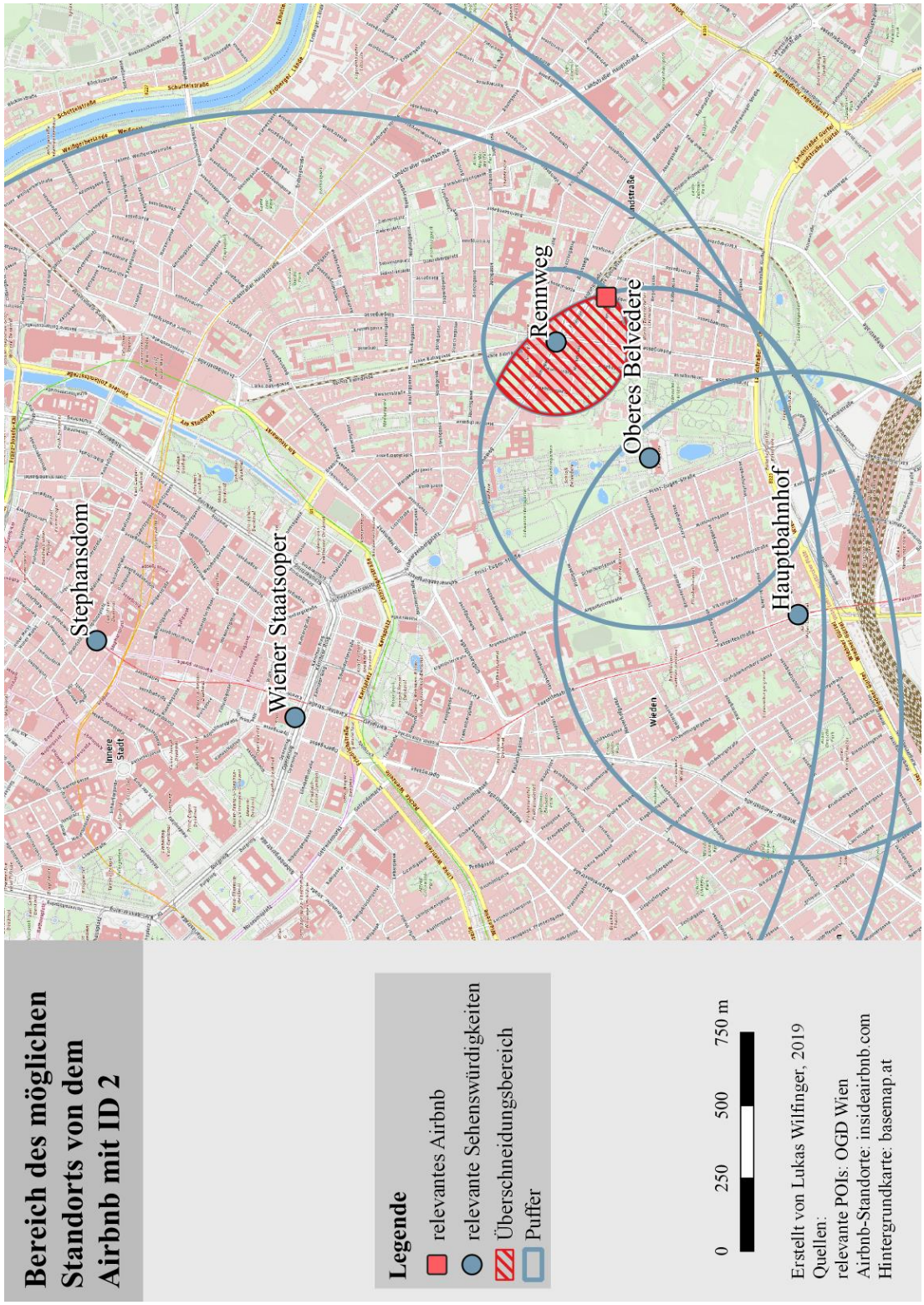
				740am to 8pm walking distance to train station nice place to stay recommend floor 1 mun walk to a supermarket which opens from 740am to 8pm walking distance to train station nice place to stay recommend this place			
10	R	Rating	1107	appartement parfait avec toutes les prestations emplacement a 50 m de la gare et 30 m du métro idéal pour visiter vienne wotjek est vraiment disponible par	3		
10	Z	Rating	1117	helpful would stay here again	0		
10	Z	Rating	1118	my family and i had a wonderful 5d4n stay in vienna and all this was made possible with our stay	3		
10	Z	Rating	1120	that make you feeling at home the location is very convenient just 5 mins from the central station and very close to tram and underground services	2	5	Hauptbahnhof
10	Z	Rating	1121	very accurate description the location is super 10 minutes to the center of wien using the metro u1 to stephenplatz	0		
10	Z	Rating	1127	woktek's place was as described extremely conveniently located his place is just 5 mins walk from the nearest station the house is well maintained and	1		
10	Z	Rating	1131	wojtek has a great apartment that is just 1 2 minutes away from the subway which makes all of vienna easily reachable clean as described great value for the money	1		
10	Z	Rating	1132	the space is a 5 min walk from the main rail station which made it very convenient	2	5	Hauptbahnhof
10	Z	Rating	1136	superschnelle kommunikation guter essenstipp genaueste beschreibungen des wohnungszugangs 1 minute zur ubahn 5 min zum stephansplatz fußgängerzone in unmittelbarer nähe alles sehr gut danke wojtek	1		
10	Z	Rating	1161	<U+AC15><U+CD94><U+D569><U+B2C8><U+B2F9>	0		
10	Z	Rating	1162	wojtek apartment is very comfortable and new the metro station only 2min away walking very easy to get around in vienna highly recommend his	1		
10	Z	Rating	1165	the apartment is about 5 minutes walk to the nearest subway exit and 10 minutes walk to the main train station there are eating outlets and a grocery outlet within a few minutes walk so it is convenient the floor is heated centrally so no issues if you have bulky luggages the host usually responded within ten minutes or so to my hidden by airbnb messages which was great	2	10	Hauptbahnhof

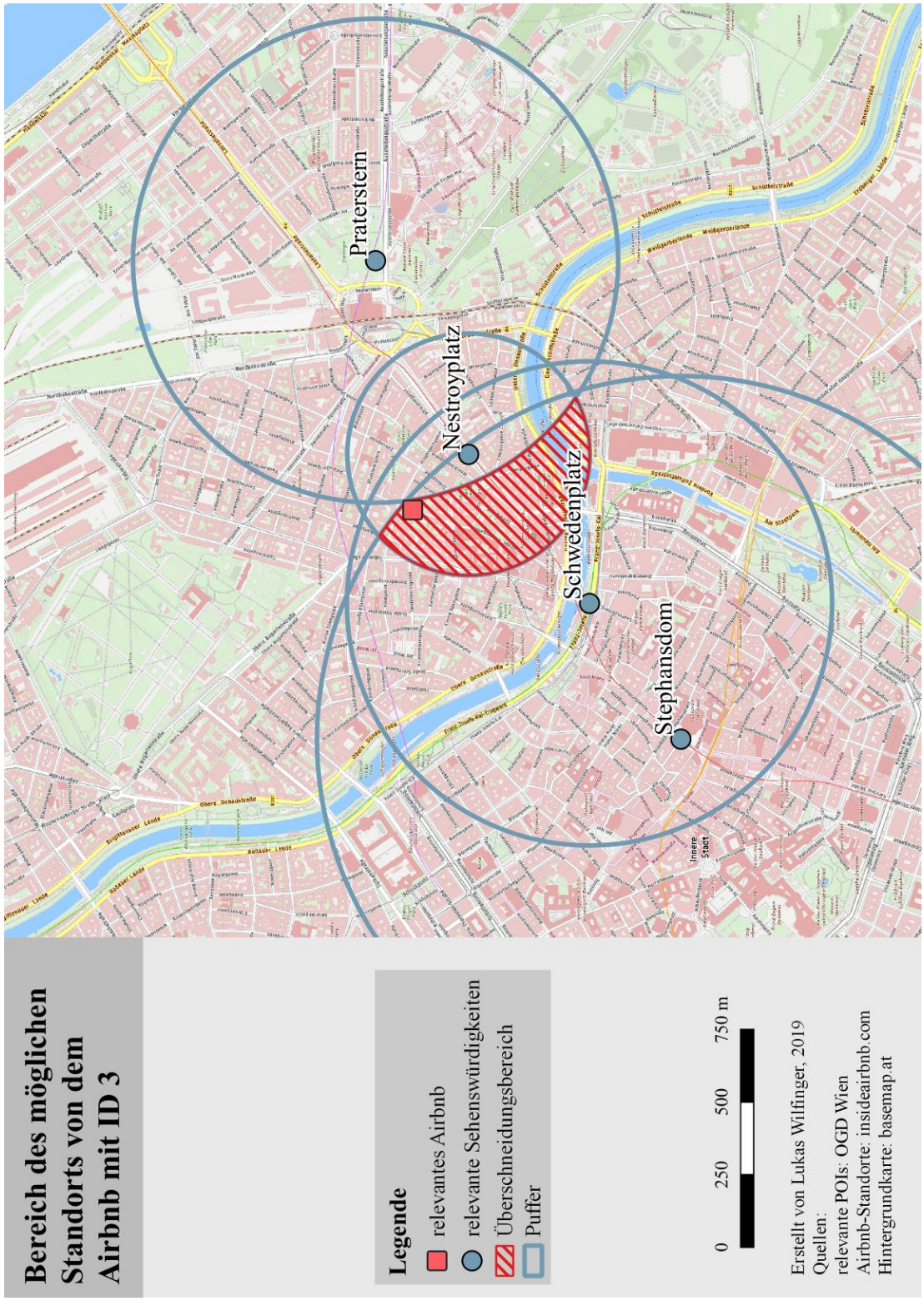
10	Z	Rating	1168	wojtek's apartment is very comfortable metro station is 3 minutes walk and if you are coming by car there is a 5 eur day parking place only 10 min walk from the flat which is great wojtek gave us great	1		
11	Z	Rating	2137	moderne calme et bien equipé il est assez éloigné du centre 20 25 minutes à pied du coeur de la ville mais il y a	3		
11	Z	Rating	2152	is close to the tram that takes you into central vienna in 15min and there is a spar supermarket 2 mins walk away i really you into central vienna in 15min and there is a spar supermarket 2 mins walk away i really enjoyed my stay it was great value	1		
11	Z	Rating	2155	fantastic only 2 mins walk to the no 2 tram station the apartment is clean	1		
11	Z	Rating	2177	kiki and mary's place was in a great location roughly 20 minutes to the old city and they were both really friendly and	1		
12	R	Host	4	gleich um die ecke die distanz zur nächsten u bahnstation beträgt etwa 500m parkplätze findet man vor der haustüre aber auch eine tiefgarage in bequer man vor der haustüre aber auch eine tiefgarage in bequer entfernung von 200m	1		
12	Z	Rating	809	a lot of character it’s relative centrical with a lot of shops 15 min walk to the city center if you wanna a coozy minimalistic	1		
12	Z	Rating	813	bezirk gelegen und wunderbar mit u bahn und tram erreichbar innerhalb von 10 min ist man am schottenring und der innenstadt die öffentlichen verkehrsmittel fahren	0		
12	Z	Rating	833	very easily by a 2 minute walk to the tram station and 10 minutes by tram we really recommend this place	1		
12	Z	Rating	835	center but right on the 40 and 41 trams which are about 10 minutes to the core would stay again	1		
13	Z	Host	8	transport the studio is a casual 10 min walk from u4 underground station hietzing the tram stop for line	2	10	Hietzing
13	Z	Rating	1185	check up on you the metro to go to city centre was 5min away overall i enjoyed my stay there and would recommend claudia's studio	1		
13	Z	Rating	1187	sono trovata molto bene in questo piccolo monolocale ha tutto e a 20 min dalla metro più vicina	3		
13	Z	Rating	1189	take the metro or the tram no 49 the journey takes about 30 mins i much preferred the tram and took it everyday one day	0		

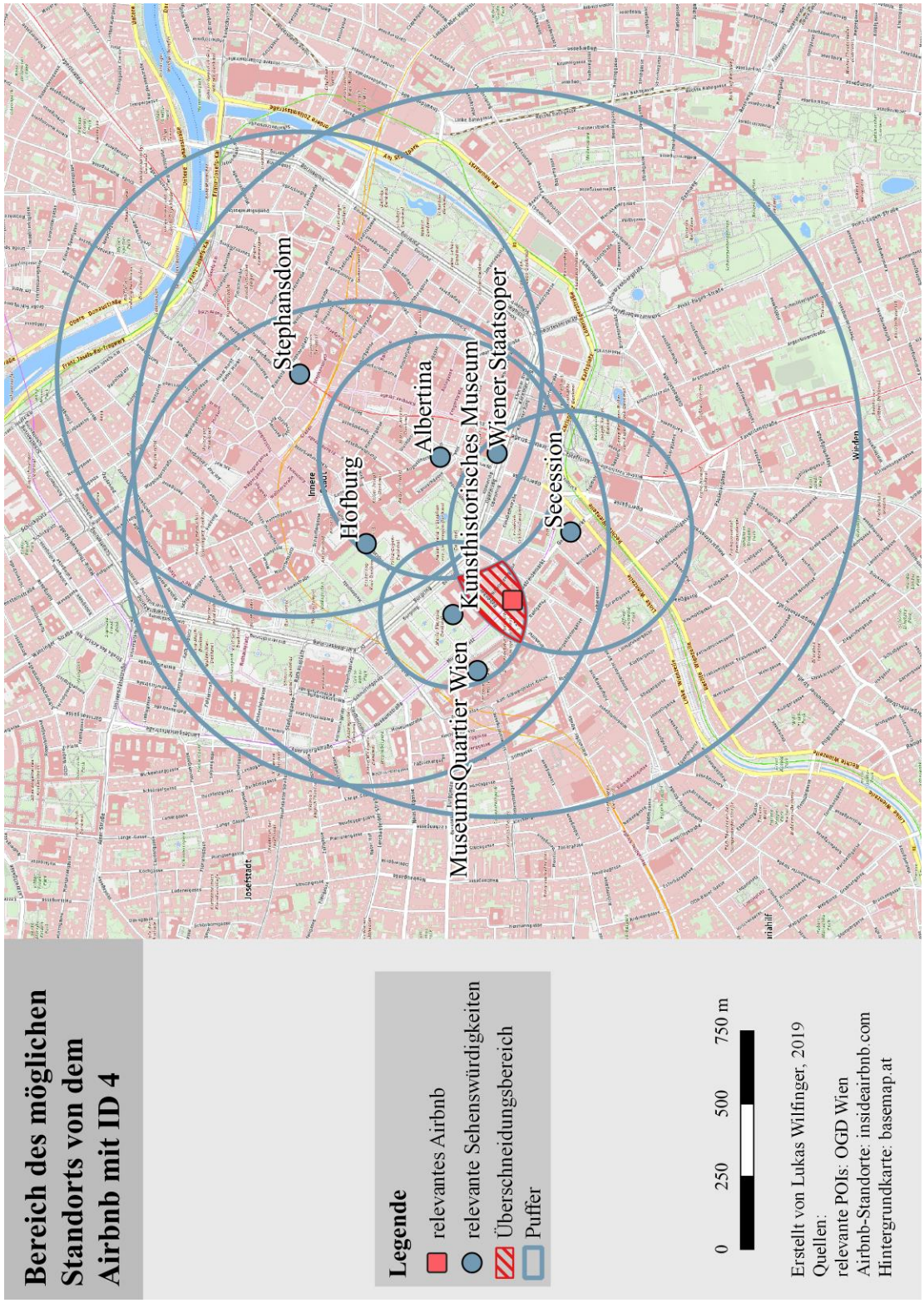
13	R	Rating	1196	were full of impressions of beautiful vienna the tram stop is only 100 m far to get city center thanks to milan for the help	1
13	Z	Rating	1197	shopping brands in a few steps reachable public transport reachable in max 10 min underground or less tramway great host and fast answer to any	0

Anhang IV Kartendarstellung der einzelnen Überschneidungsbereiche für jedes Airbnb und nach Puffer- oder Isochronen-Distanzen

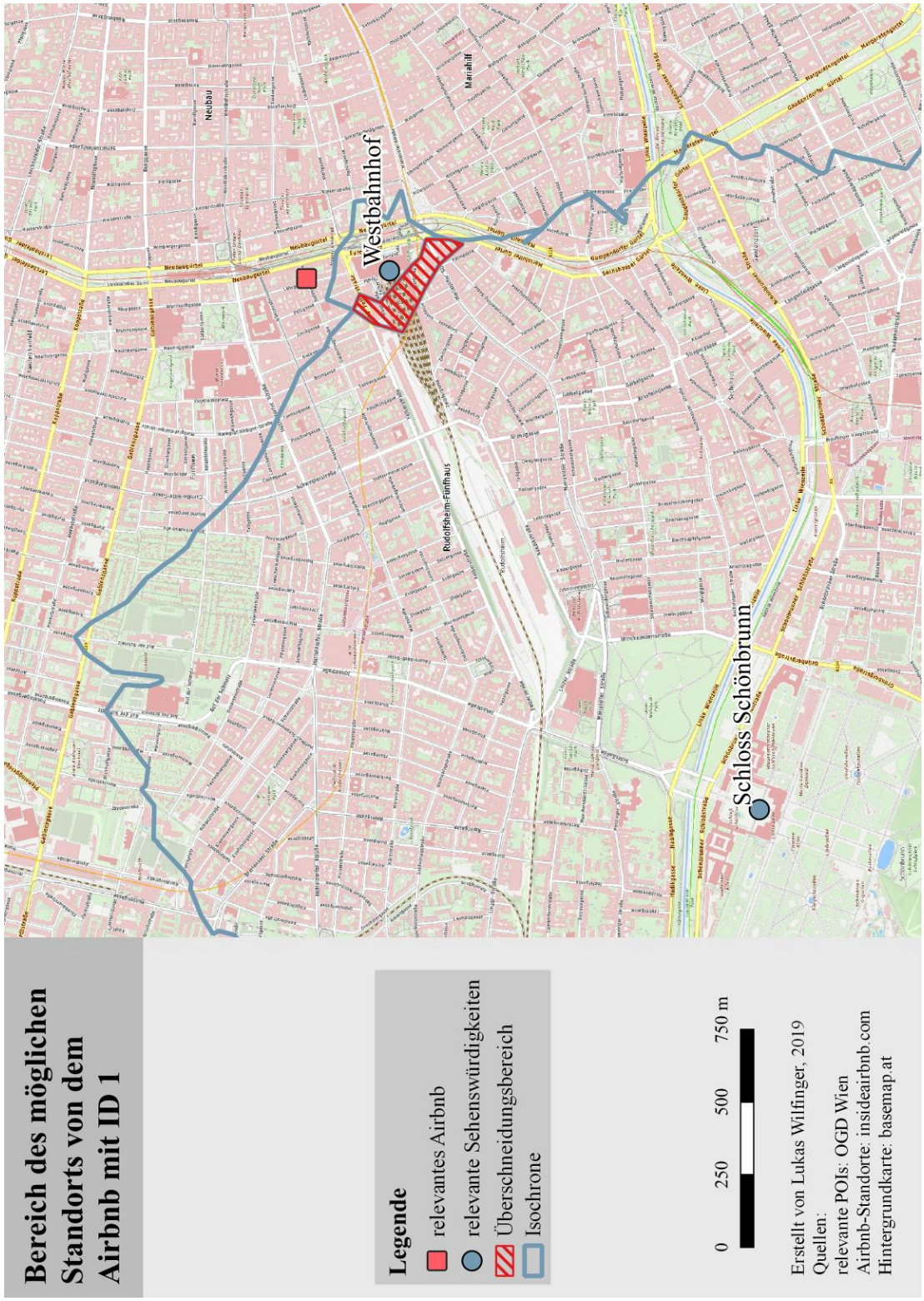


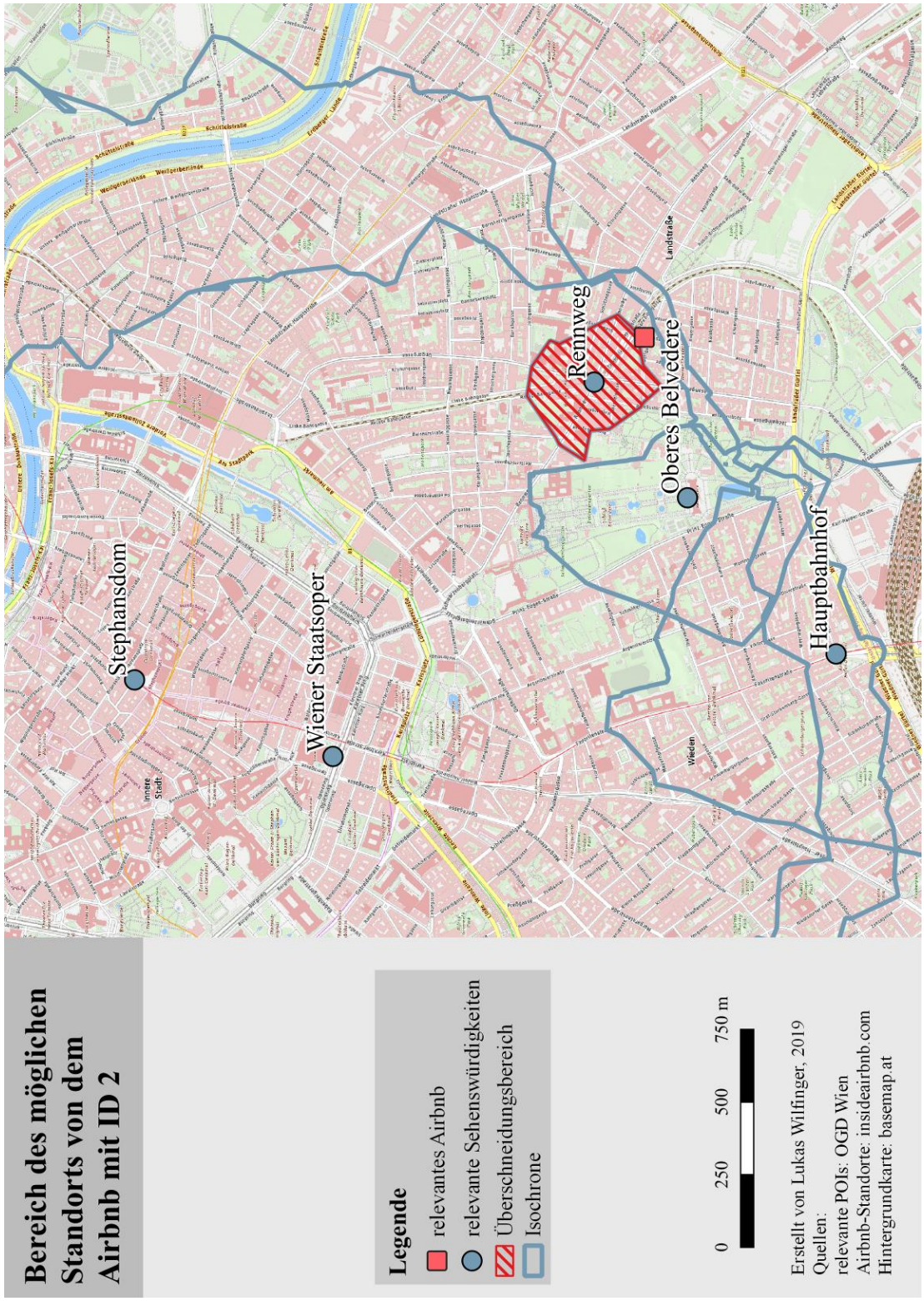

Bereich des möglichen Standorts von dem Airbnb mit ID 2
Legende
relevantes Airbnb
relevante Sehenswürdigkeiten
Überschneidungsbereich
Puffer
0 250 500 750 m
Erstellt von Lukas Wilfinger, 2019
Quellen:
relevante POIs: OGD Wien
Airbnb-Standorte: insideairbnb.com
Hintergrundkarte: basemap.at
Stephansdom
Wiener Staatsoper
Rennweg
Oberes Belvedere
Hauptbahnhof

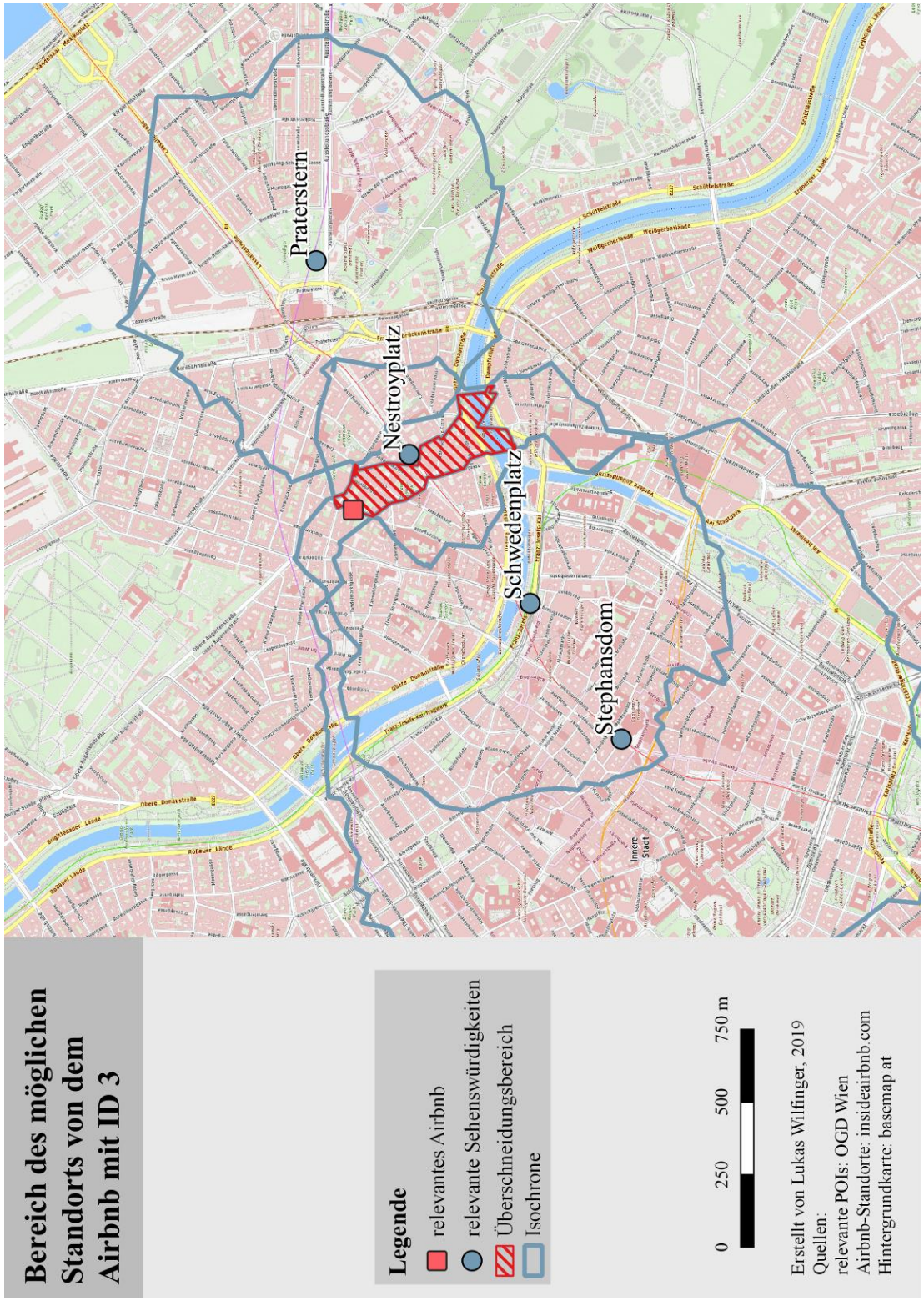

Bereich des möglichen Standorts von dem Airbnb mit ID 3
Praterstern
Nestroyplatz
Schwedenplatz
Stephansdom
Legende
relevantes Airbnb
relevante Sehenswürdigkeiten
Überschneidungsbereich
Puffer
0
250
500
750 m
Erstellt von Lukas Wilfinger, 2019
Quellen:
relevante POIs: OGD Wien
Airbnb-Standorte: insideairbnb.com
Hintergrundkarte: basemap.at

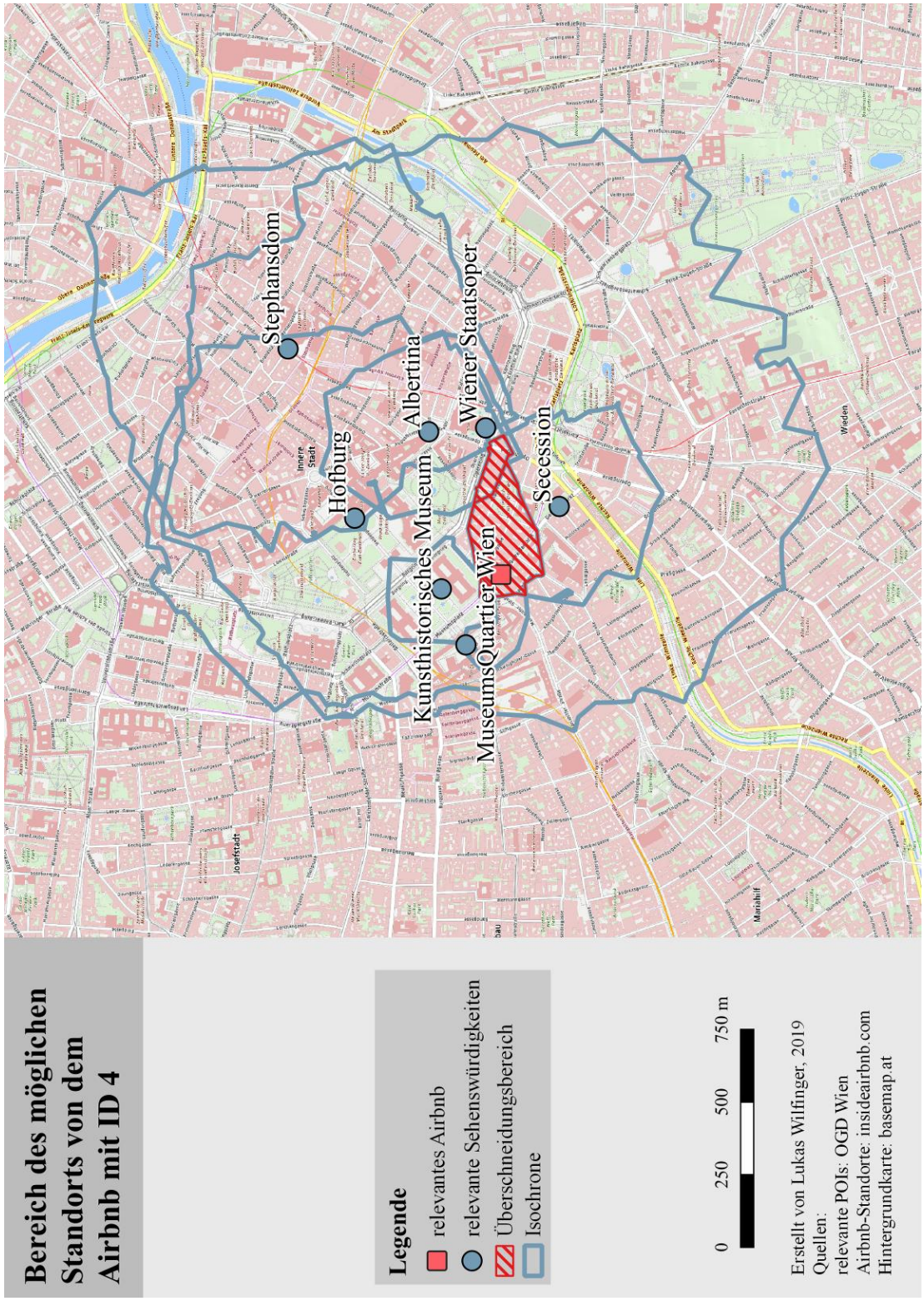

Bereich des möglichen Standorts von dem Airbnb mit ID 4
Stephansdom
Albertina
Wiener Staatsoper
Hofburg
Secession
Kunsthistorisches Museum
MuseumsQuartier Wien
Legende
relevantes Airbnb
relevante Sehenswürdigkeiten
Überschneidungsbereich
Puffer
0
250
500
750 m
Erstellt von Lukas Wilfinger, 2019
Quellen:
relevante POIs: OGD Wien
Airbnb-Standorte: insideairbnb.com
Hintergrundkarte: basemap.at

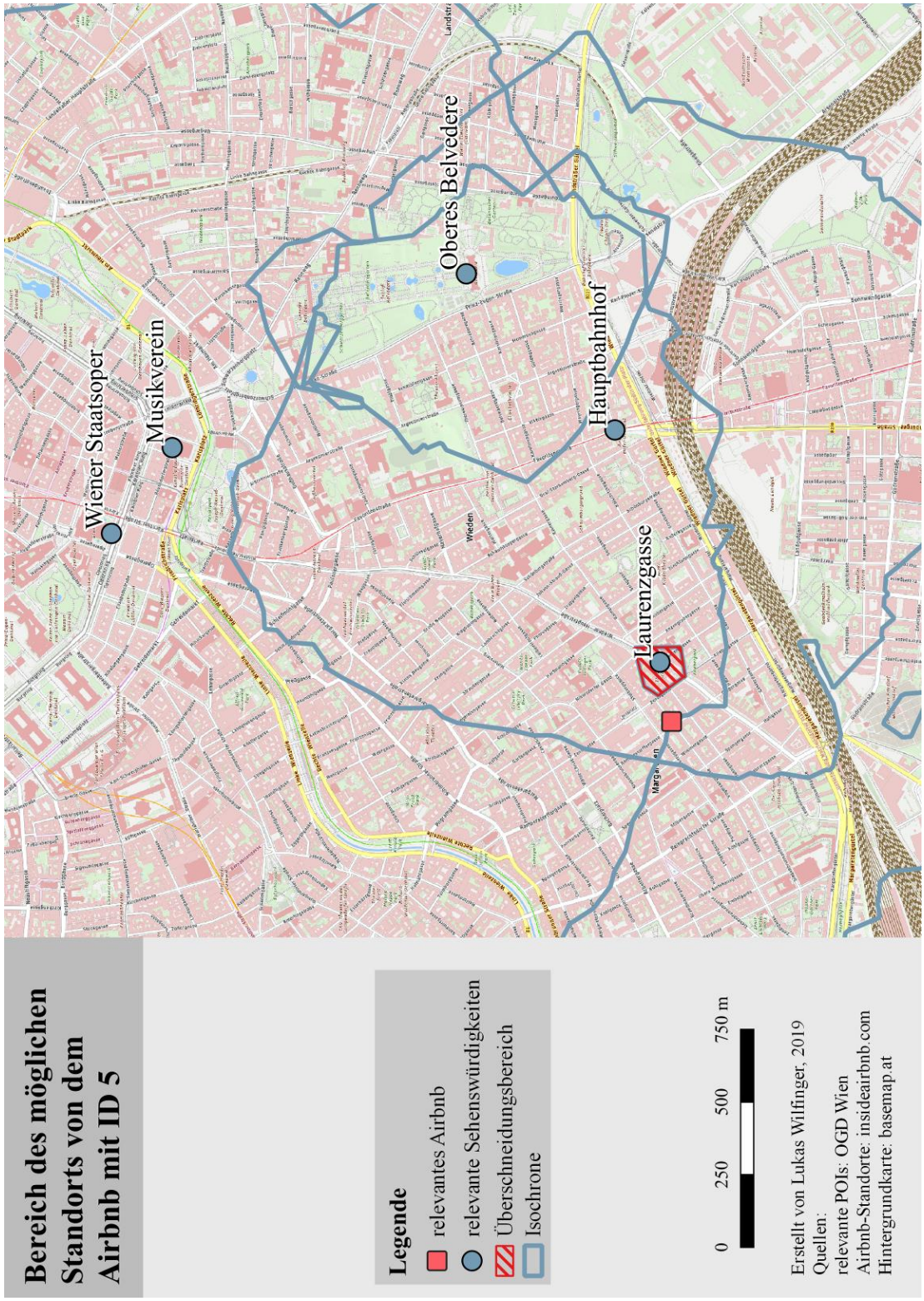
Bereich des möglichen Standorts von dem Airbnb mit ID 5
Wiener Staatsoper
Musikverein
Oberes Belvedere
Hauptbahnhof
Laurenzgasse
Margareten
Legende
relevantes Airbnb
relevante Sehenswürdigkeiten
Überschneidungsbereich
Puffer
0
250
500
750 m
Erstellt von Lukas Wilfinger, 2019
Quellen:
relevante POIs: OGD Wien
Airbnb-Standorte: insideairbnb.com
Hintergrundkarte: basemap.at

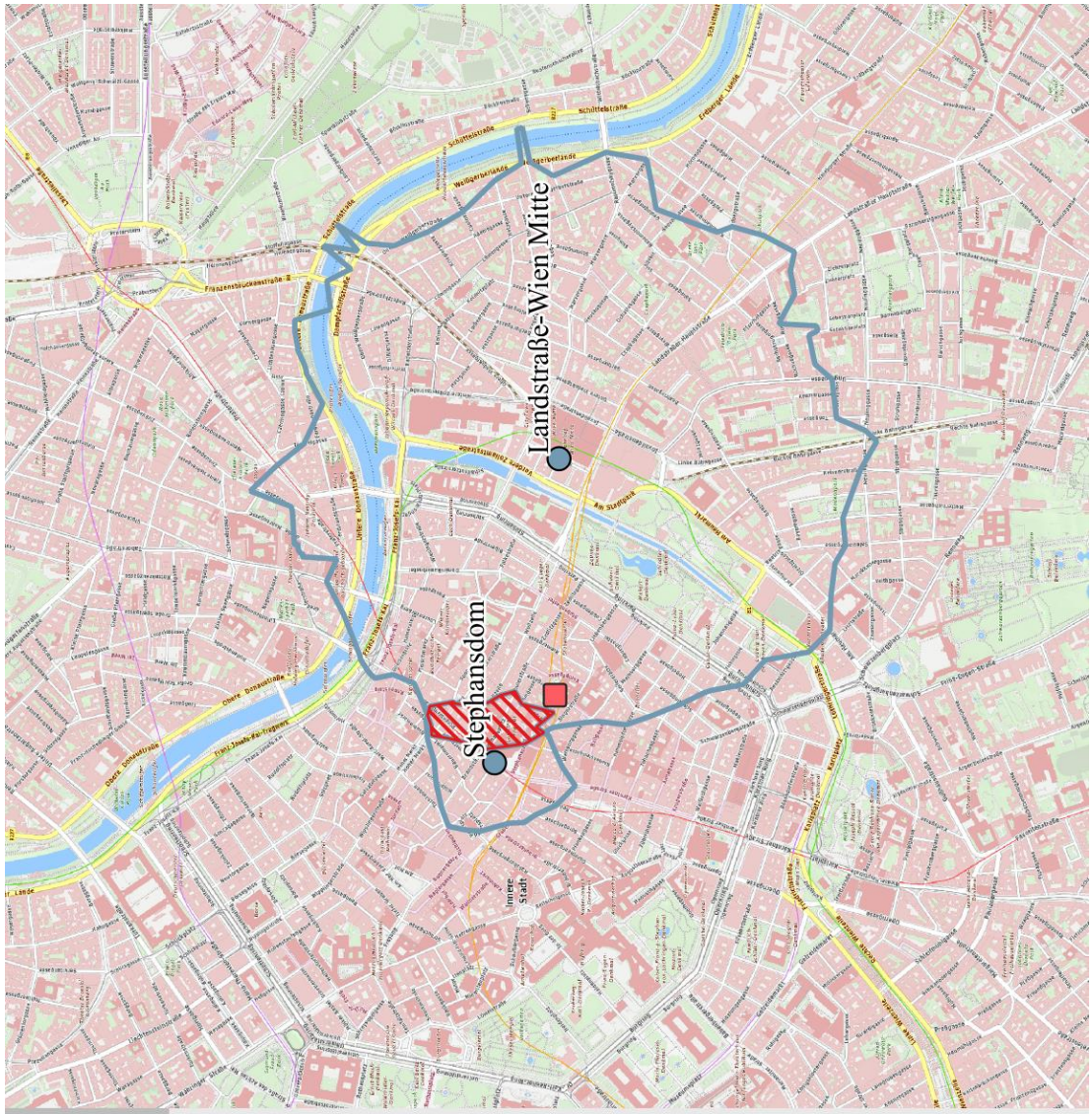

Bereich des möglichen Standorts von dem Airbnb mit ID 1
Legende
relevantes Airbnb
relevante Sehenswürdigkeiten
Überschneidungsbereich
Isochrone
0 250 500 750 m
Erstellt von Lukas Wilfinger, 2019
Quellen:
relevante POIs: OGD Wien
Airbnb-Standorte: insideairbnb.com
Hintergrundkarte: basemap.at
Westbahnhof
Schloss Schönbrunn
Rudolfsheim-Fünfhaus
Neubau
Mariahilf


Bereich des möglichen Standorts von dem Airbnb mit ID 2
Legende
relevantes Airbnb
relevante Sehenswürdigkeiten
Überschneidungsbereich
Isochrone
0 250 500 750 m
Erstellt von Lukas Wilfinger, 2019
Quellen:
relevante POIs: OGD Wien
Airbnb-Standorte: insideairbnb.com
Hintergrundkarte: basemap.at
Stephansdom
Wiener Staatsoper
Rennweg
Oberes Belvedere
Hauptbahnhof


Bereich des möglichen Standorts von dem Airbnb mit ID 3
Legende
relevantes Airbnb
relevante Sehenswürdigkeiten
Überschneidungsbereich
Isochrone
0 250 500 750 m
Erstellt von Lukas Wilfinger, 2019
Quellen:
relevante POIs: OGD Wien
Airbnb-Standorte: insideairbnb.com
Hintergrundkarte: basemap.at
Praterstern
Nestroyplatz
Schwedenplatz
Stephansdom


Bereich des möglichen Standorts von dem Airbnb mit ID 4
Legende
relevantes Airbnb
relevante Sehenswürdigkeiten
Überschneidungsbereich
Isochrone
0 250 500 750 m
Erstellt von Lukas Wilfinger, 2019
Quellen:
relevante POIs: OGD Wien
Airbnb-Standorte: insideairbnb.com
Hintergrundkarte: basemap.at
Stephansdom
Hofburg
Albertina
Wiener Staatsoper
Secession
Kunsthistorisches Museum
MuseumsQuartier Wien
Innere Stadt
Josefstadt
Wieden
Mariahilf


Bereich des möglichen Standorts von dem Airbnb mit ID 5
Wiener Staatsoper
Musikverein
Oberes Belvedere
Hauptbahnhof
Laurenzgasse
Wieden
Legende
relevantes Airbnb
relevante Sehenswürdigkeiten
Überschneidungsbereich
Isochrone
0
250
500
750 m
Erstellt von Lukas Wilfinger, 2019
Quellen:
relevante POIs: OGD Wien
Airbnb-Standorte: insideairbnb.com
Hintergrundkarte: basemap.at



Bereich des möglichen Standorts von dem Airbnb mit ID 7

Legende

- relevantes Airbnb
- relevante Sehenswürdigkeiten
- Überschneidungsbereich
- Isochrone

0 250 500 750 m

Erstellt von Lukas Wilfinger, 2019
Quellen:
relevante POIs: OGD Wien
Airbnb-Standorte: insideairbnb.com
Hintergrundkarte: basemap.at

Lebenslauf

Bildungsweg

seit 03/2016	Masterstudium Kartographie und Geoinformation Universität Wien
2016 - 2018	Masterstudium Data Science FH Wiener Neustadt (am 21.06.2018 mit Auszeichnung abgeschlossen)
2013 - 2016	Bachelorstudium Geographie Universität Wien (am 06.04.2016 mit einem Notenschnitt von 2,16 abgeschlossen)
2006 - 2012	HTL St.Pölten Schwerpunkt: Technische Informatik
2002 - 2006	BRG Lilifenfeld
1998 - 2002	Volksschule Hainfeld

Berufserfahrung

seit 08/2018	Data Scientist Österreichische Post AG
02/2018 – 05/2018	Data Scientist – Master Thesis Traffic Consultants GmbH
08/2017 – 09/2017	Praktikum Traffic Consultants GmbH
07/2017	Praktikum Bundesforschungs- und Ausbildungszentrum für Wald, Naturgefahren und Landschaft
09/2015	Praktikum GeoMarketing & GIS Solutions und Services GmbH
08/2015	Praktikum Land-, forst- und wasserwirtschaftliches Rechenzentrum GmbH
09/2012 – 03/2013	Präsenzdienst beim Österreichischen Bundesheer in St. Pölten
Sommermonate 2008 - 2016	Prefa Aluminiumprodukte, Neuman Aluminium GmbH, Voest Alpine Giesserei, Gemeinde Rohrbach und Schmid Schrauben GmbH

Ich versichere:

- dass ich die Masterarbeit selbstständig verfasst, andere als die angegebenen Quellen und Hilfsmittel nicht benutzt und mich auch sonst keiner unerlaubten Hilfe bedient habe.
- dass alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten und nicht veröffentlichten Publikationen entnommen sind, als solche kenntlich gemacht sind.
- dass ich dieses Masterarbeitsthema bisher weder im In- noch im Ausland (einer Beurteilerin/ einem Beurteiler zur Begutachtung) in irgendeiner Form als Prüfungsarbeit vorgelegt habe.
- dass diese Arbeit mit der vom Begutachter beurteilten Arbeit übereinstimmt.

Datum Unterschrift