



UNIVERSITÄT  
LEIPZIG

Methods bazaar 2023 talk

# The travels of Marco Polo

Information extraction and visualization of historic  
travel literature

14.02.2023

Andreas Niekler, Magdalena Wolska, Marvin Thiel, Yiwen Cao

Computational  
Humanities

UNIVERSITÄT LEIPZIG

# COVERED MATERIAL

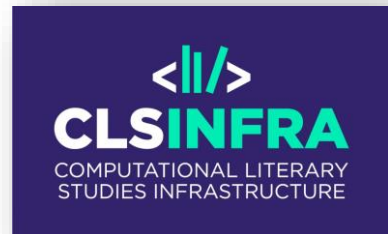
## OUTLINE

- **Introduction**
  - **Computational Literature Studies**
  - **Travel Literature**
  - **Marco Polo**
  - **Mining Travel Literature**
- **Natural Language Processing**
  - **Corpus**
  - **Named Entity Recognition**
  - **Motion Identification**
  - **Tools**
  - **Evaluation**
- **Spatial Analysis and Visualization**
- **Future Directions**

# (COMPUTATIONAL) LITERARY STUDIES

## THE STUDY OF LITERATURE WITH COMPUTATIONAL METHODS

- Application of **data science, computer science, and close reading**
- **Computational methods** for the analysis of literary texts and their (cultural, social, historical, performative) contexts
- Methods include:
  - Genre classification
  - Stylometry
  - Attribution of characters
  - Speech, Thoughts and Writing Analysis
  - Relation, Networks
  - Modelling of Narration
  - Segmentation, Structuring

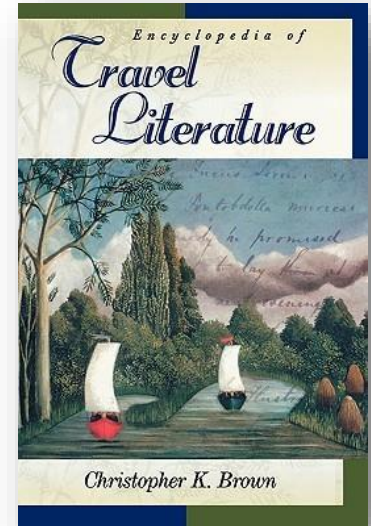


[1] Priority programme SPP 2207 Computational Literary Studies (CLS), funded by the German Research Foundation (DFG)

# TRAVEL LITERATURE

## SUBTEXT OF A LOT OF LITERATURE [2]

- Travel is the **basic underpinning or subtext of a lot of literature** [2]
- The genre of travel literature encompasses **outdoor literature, guidebooks, nature writing, and travel memoirs**. [3]
- Travel literature has its **roots in this reality** [3]
- A human being develops itself by confronting itself with reality and the “unknown” [4]
- Many examples: Che Guevara’s *The Motorcycle Diaries*, *The Travels of Marco Polo (Il Milione)*, Peter Mayle’s *A Year in Provence*, ...



Source: Goodreads

[2] Mewshaw, M. (2005). Travel, Travel Writing, and the Literature of Travel. *South Central Review*, 22(2), 2–10.  
<https://doi.org/10.1353/scr.2005.0042>

[3] Cuddon, J. A., & Birchwood, M. (2014). *The Penguin dictionary of literary terms and literary theory* (R. Habib, Hrsg.; 5. ed., publ. in paperback). Penguin Books.

[4] Brenner, Peter: *Does Travelling Matter? The Impact of Travel Literature on European Culture*.

# MARCO POLO

## VENETIAN MERCHANT AND EXPLORER

- **Traveled through Asia along the Silk Road** between 1271 and 1295. [4]
- *Il milione* (“The Million”), **known in English as the *Travels of Marco Polo***, is a classic of travel literature. [4]
- 13th-century travelogue written down by Rustichello da Pisa from stories told by Italian explorer Marco Polo. [5]
- The book was translated into many European languages in Marco Polo's own lifetime. [5]



Source: [5]

[5] <https://www.britannica.com/biography/Marco-Polo>

[6] [https://en.wikipedia.org/wiki/The\\_Travels\\_of\\_Marco\\_Polo](https://en.wikipedia.org/wiki/The_Travels_of_Marco_Polo)

# MARCO POLO'S TRAVELOGUE

TO WHAT EXTEND REAL? CAN WE TELL?

- Some have questioned **whether Marco had actually travelled to China or was just repeating stories** that he had heard from other travellers:

“Countless authors of travelogues, such as Marco Polo, presented often rather astonishing accounts seemingly unbelievable in their content for their audiences back home” [7, p. 27]



Source: [6]

[7] Classen, A. (2013). *East Meets West in the Middle Ages and Early Modern Times: Transcultural Experiences in the Premodern World*. De Gruyter.

# MARCO POLO'S TRAVELOGUE



Source: [5]

# MINING TRAVEL LITERATURE

## OUR WORK

- **Basic exploratory work on the genre of travel literature**
  - Can we **reconstruct the route** of Marco Polo by analyzing the text contents **semi-automatically**?
  - Is it possible to **geo-reference location entities** from the book?
- **Visualization of travel** and movement
  - Can we use the information gained to **create immersive visualizations** to **augment the reading experience**?
  - Can visualizations be used to **align readers experience and reality**?



# MINING TRAVEL LITERATURE

## OUR WORK

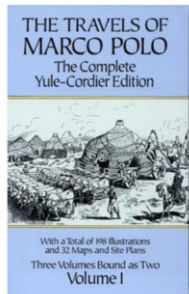
- **Step 1: Natural Language Processing/Information Extraction**
- **Step 2: Geo-Referencing of entities**
- **Step 3: Routes and partial routes visualization**

# NATURAL LANGUAGE PROCESSING








## CORPUS

- English translation, Text file, **Sentence segmentation processing**, Translation by Henry Yule, Published in 2 volumes
  - <https://www.gutenberg.org/ebooks/10636>,
  - <https://www.gutenberg.org/ebooks/12410>

### The Travels of Marco Polo — Volume 1 by Marco Polo and Rustichello of Pisa



#### Download This eBook

Format 	Size
 <a href="#">Read this book online: HTML</a>	3.1 MB
 <a href="#">EPUB (with images)</a>	27.6 MB
 <a href="#">EPUB (no images)</a>	1.2 MB
 <a href="#">Kindle (with images)</a>	28.4 MB
 <a href="#">Kindle (no images)</a>	2.0 MB
 <a href="#">Plain Text UTF-8</a>	2.2 MB

# NATURAL LANGUAGE PROCESSING

## CORPUS

<b>Part</b>	<b>Description</b>	<b>Total Section</b>
Prologue	It describes briefly about the journey of the Polos'.	18
Book I	It describes the journey from the Lesser Armenia to the Court of the Great Kaan at Chandu.	61
Book II	It describes the Great Kaan, his capital city and the customs in Cathy and the journey through the Cathy and Manzi.	82
Book III	It describes Japan, the Archipelago, Southern India and the Coasts and Islands of the Indian Sea.	40
Book IV	It describes the wars among the Tartar Lords, and the Northern Countries.	34

# NATURAL LANGUAGE PROCESSING

## TOOLS

- NLP
  - Flair: State-of-the-art **Natural Language Processing** framework based on Pytorch and built on Python (<https://github.com/flairNLP/flair>)
  - NLTK: powerful platform to do **Natural Language Processing jobs in Python** (<https://www.nltk.org/>)
  - Stanford Parser: **constituency parsing** used shift-and reduce operations (<https://stanfordnlp.github.io/CoreNLP/>)
- Lexical Resources
  - VerbNet: largest online English **verb lexicon** (<https://verbs.colorado.edu/verbnet/>); includes verbs' predicate-argument structures
  - FrameNet: annotated **examples of words' meaning** and usage in real texts (<https://framenet.icsi.berkeley.edu/fndrupal/>); frame annotated
  - Semlink: **mapping files to link different lexical resources** together and link their annotated instances (<https://github.com/cu-clear/semlink>)
  - WordNet: large **lexical database** of English (<https://wordnet.princeton.edu/>)

# NATURAL LANGUAGE PROCESSING

## TOOLS

- GIS Systems
  - GeoNames: **geographical database covers all countries** and contains over 25,000,000 placenames of different language (<https://www.geonames.org/>)
  - CHGIS: **China Historical Geographic Information System** (<https://chgis.fas.harvard.edu/>)
  - SRHGIS: stands for **Silk Road Historical GIS** (<https://www.srhgis.com/>)
  - SRGIS: Silk Road GIS (<http://silkroad.fudan.edu.cn/project.html>)

# NATURAL LANGUAGE PROCESSING

## ANNOTATION

- For the gold standard **Named Entity Recognition annotations**, the clear boundaries and the types of the named entities are important.
  - Most of the observed named entities contain one or two tokens, still some contain more than three tokens.
  - Named Entity Recognition model from Flair was used to identify the Named Entities in the sentence.
  - The difference between the annotations from Flair model and the manual annotations is the performance of this model.
- The second gold standard annotations are about the **motion verbs**.
  - With the help of the **lexical resources, a list of possible motion verbs is built**.
  - These verbs are then marked in the sentences automatically and tested
- The third gold standard annotations are about the **motion events**.
  - Part-of-Speech tag and parser, **noun phrases and prepositional phrases belonging to the motion verb are identified**.

# NATURAL LANGUAGE PROCESSING

## GAZETTEER

- **Ambiguity resolution**
- Entity **referencing, linking**
- The building of the gazetteer for this work is **based on the index from the back matter** of “The Travels of Marco Polo”
- Both **indexes from Henry Yule and Hugh Murray**, the translators and editors, are used in this thesis.

Entity Name	Elobrations	Related Contents	Alternative Name	Type
Kinsay	[‘formerly Lin-ngan now Hang-chau fu’]	[‘its surrender to Bayan;’, ‘extreme public security;’, ‘...]	[‘Capital’, ‘ Khansa’, ‘ Khinsa’, ‘ Khingsai’, ‘ Khanzai’, ‘Cansay’, ‘Campsay’, ‘Kin-sai’, ‘Quin-sai’, ‘Kitvaal’, ‘Quin-sal’]	GPE

# NATURAL LANGUAGE PROCESSING

## NER AND RESOLVING AMBIGUITIES IN HISTORICAL CONTEXT

- **Modern names differ** from historic names
- Existing **models don't satisfy** in performance
- Models must be **adopted and re-trained**

**Table 1.1:** Examples of the Location names used by Marco Polo and the Modern Names

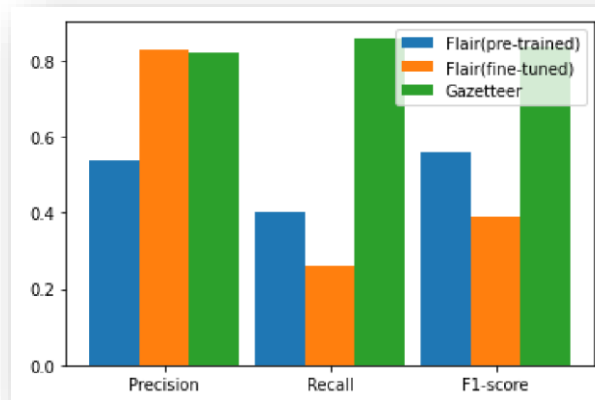
<b>Polo's Name</b>	<b>Modern Name</b>
Paipurth	Bayburt
Cacanfu	Hejianfu
Sinju	Xining



# NATURAL LANGUAGE PROCESSING

## NER PERFORMANCE

- Named Entity Recognition model from **Flair** was used to identify the **Named Entities** in the sentence.
- The **difference between the annotations from Flair model and the manual annotations** is the performance of this model.
- **Flair was fine-tuned based on the gold standard annotations**



# NATURAL LANGUAGE PROCESSING

## MOTION EVENTS IDENTIFICATION

- Locations in the travel route are **linked with travel verbs**
- Finding verbs which express motion, can be a **fast way to identify whether the locations are in the travel routes**
  - The ideal situation for extracting a motion event is to get a clear origin and destination:  
*...that you pass in **going** from Trebizond to Tauris...From Tauris to Persia is a journey of twelve days."*
- Direction between the two locations and the length of the dates travelled are also specified
  - ***Travelling** through a succession of towns and villages that look like one continuous city, two days further on to the south-east, you find the great and fine city of **GHIUJU** which is under **Kinsay**."*

# NATURAL LANGUAGE PROCESSING

## MOTION EVENTS IDENTIFICATION

- Location and Motion Linking in **non-chronological Context**
- The book does not only contain the **motion events** whose agent is Marco Polo's, but also **other characters** in the stories
  - Motion Verbs of other Characters:  
“Chinghis Kaan with all his host *arrived* at a vast and beautiful plain which was called Tanduc.”
  - Motion Verbs of Marco Polo:  
“When you have *gone* 15 miles from the city of Unken, you *come to* this noble city which is the capital of the kingdom.”
- By **identifying the subject of the motion verb**, we tried to classify whether the **location name attached to the moving verb** is in Marco Polo's route

# NATURAL LANGUAGE PROCESSING

## MOTION EVENT EXTRACTION

- We use both **VerbNet** and **FrameNet** to generate the motion verb list
- In VerbNet, the **verbs classes** under a top-level number share a **similar semantic meaning**.
- Class number 51 represents '**Verbs of Motion**'.
- By checking the semantics of the verb classes, we choose *'reach-51.8', 'meander-47.7', 'escape-51.1-2-1', 'escape-51.1-1', 'escape-51.1-2', 'leave-51.2-1', 'roll-51.3.1', 'run-51.3.2', 'nonvehicle-51.4.2'* as the subclasses

# NATURAL LANGUAGE PROCESSING

## MOTION EVENT EXTRACTION

No Comments

run-51.3.2-2

Members: 31, Frames: 5

### MEMBERS

BACK (FN 1; WN 2)	FLOAT (FN 1; WN 1, 2; G 1)	LOPE (FN 1; WN 1)
BOUNCE (WN 1, 2, 3; G 1, 6)	GALLOP (WN 1, 2, 3; G 1)	MARCH (WN 2, 6; G 2)
BOWL (WN 1; G 3)	GLIDE (FN 1; WN 1, 3; G 1)	PRANCE (FN 1; WN 1)
CANTER (FN 1; WN 1, 2, 3; G 1)	HASTEN (FN 1; WN 1, 2; G 1)	RACE (WN 1, 4; G 1)
COAST (FN 1; WN 1)	HOBBLE (FN 1; WN 1; G 1)	ROLL (FN 1, 2, 3, 4; WN 1, 9; G 1)
DART (FN 1; WN 1, 2; G 1)	HURRY (FN 1; WN 1; G 1)	SCOOT (FN 1; WN 1)
DASH (FN 1; WN 1; G 1)	HURTLE (WN 1, 2, 3)	SKIP (FN 1; WN 3, 6; G 2)
DRIFT (FN 1; WN 1, 2, 6, 7, 8; G 1, 2, 4)	INCH (WN 1; G 1)	SKITTER (WN 1, 2, 3; G 1)

### ROLES

- RESULT

### FRAMES

#### NP V NP PP.LOCATION

EXAMPLE "Tom jumped the horse over the fence."

SYNTAX AGENT V THEME {{+SPATIAL}} LOCATION

SEMANTICS MOTION(DURING(E0), THEME) PREP(E0, THEME, LOCATION) CAUSE(AGENT, E0) EQUALS(E0, E1) MOTION(DURING(E1), AGENT) PREP(E1, AGENT, LOCATION)

#### NP V NP PP.LOCATION

EXAMPLE "The lion tamer jumped the lions through the loop."

SYNTAX AGENT V THEME {{+SPATIAL}} LOCATION

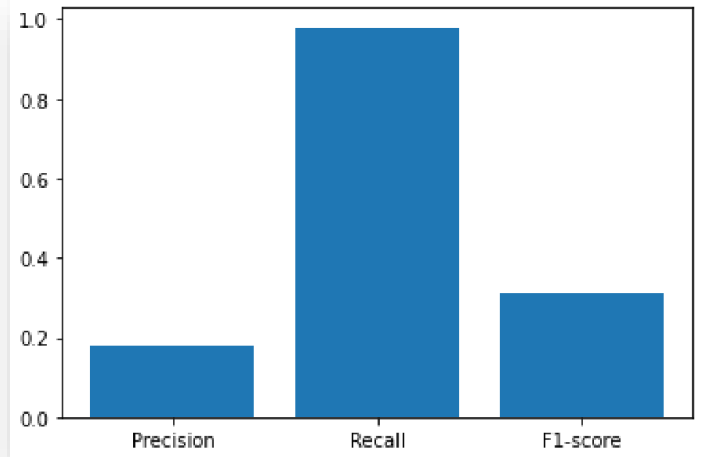
SEMANTICS MOTION(DURING(E), THEME) PREP(E, THEME, LOCATION) CAUSE(AGENT, E)

<https://verbs.colorado.edu/verb-index/vn/run-51.3.2.php>

# NATURAL LANGUAGE PROCESSING

## MOTION EVENT EXTRACTION

- We divide the process of finding motion verbs into **three steps**:
  - **Generating motion verb list** from lexical resources;
  - **Marking the verbs in the sentences**, and finding the base forms of the verbs;
  - Comparing the **base forms of the verbs to the motion verb list** to identify the motion verbs.



# NATURAL LANGUAGE PROCESSING

## ROUTE EXTRACTION

- After route extraction, **398 short sentences, which express motion of Marco Polo** and contain relative exact sources, destinations or locations, are found.
- We put **133 locations on the Map** and link them with possible chronological order.
- The Result is **saved as a kml file**.



# NATURAL LANGUAGE PROCESSING

## ROUTE EXTRACTION





# SPATIAL ANALYSIS

## EXAMPLE KML

- Input to further analysis
- **KML: Keyhole Markup Language**

```
<Placemark id="016A3E7BDA20D9C8D690">
  <name>PIANFU</name>
  <gx:mid>/m/05xtzn</gx:mid>
  <gx:fid>0x3676dd556edbe6e9:0x12fdd98581592320</gx:fid>
  <LookAt>
    <longitude>111.5133548</longitude>
    <latitude>36.0938897</latitude>
    <altitude>0</altitude>
    <heading>0</heading>
    <tilt>0</tilt>
    <gx:fovy>35</gx:fovy>
    <range>37569.62472024467</range>
    <altitudeMode>relativeToGround</altitudeMode>
  </LookAt>
  <styleUrl>#__managed_style_094D212AB720D9C6CA15</styleUrl>
  <Point>
    <altitudeMode>relativeToGround</altitudeMode>
    <coordinates>111.51962,36.0882199,0</coordinates>
  </Point>
</Placemark>
<Placemark id="016A3E7BDA20D9C8D690">
  <name>CACHANFU</name>
  <gx:mid>/g/15dphsrd</gx:mid>
  <gx:fid>0x367b007648e7ec71:0x1c4e6568a597f818</gx:fid>
  <LookAt>
    <longitude>110.32919</longitude>
    <latitude>34.8341595</latitude>
    <altitude>0</altitude>
    <heading>0</heading>
    <tilt>0</tilt>
    <gx:fovy>35</gx:fovy>
    <range>4418.384224441332</range>
    <altitudeMode>relativeToGround</altitudeMode>
  </LookAt>
  <styleUrl>#__managed_style_094D212AB720D9C6CA15</styleUrl>
  <Point>
    <altitudeMode>relativeToGround</altitudeMode>
    <coordinates>110.32919,34.83416,0</coordinates>
  </Point>
</Placemark>
```

# SPATIAL ANALYSIS

## RENDERING MAPS INSTEAD OF GOOGLE MAPS

- There is **more to travel experiences than just the places** visited alone.
- Vegetation, climate, obstacles and vistas are all **details that come to mind when thinking about traveling**, especially by foot.
- Details about **landscape and the traveled roads are most often sparsely described** and days worth of travel are described using just a few sentences

“When you *depart from* this City of Cobinan, you find yourself again in a Desert of surpassing aridity, which lasts for some eight days; here are neither fruits nor trees to be seen, and what water there is bitter and bad, so that you have to carry both food and water.”

# SPATIAL ANALYSIS

## RENDERING MAPS INSTEAD OF GOOGLE MAPS

- However, when **analyzing travel writing, regular point-based maps do not normally allow for analyzing aspects of a more qualitative nature** like the experiences that were made while traveling [8].
- [9] describe an environment with the purpose "**to create immersive geographies that link the experiential, the emotional and the symbolic elements of literary works to the nuanced, dimensional richness of places as inspired by authors and their works**".

[8] Murrieta-Flores et al., 2016

[9] Trevor M. Harris et al.

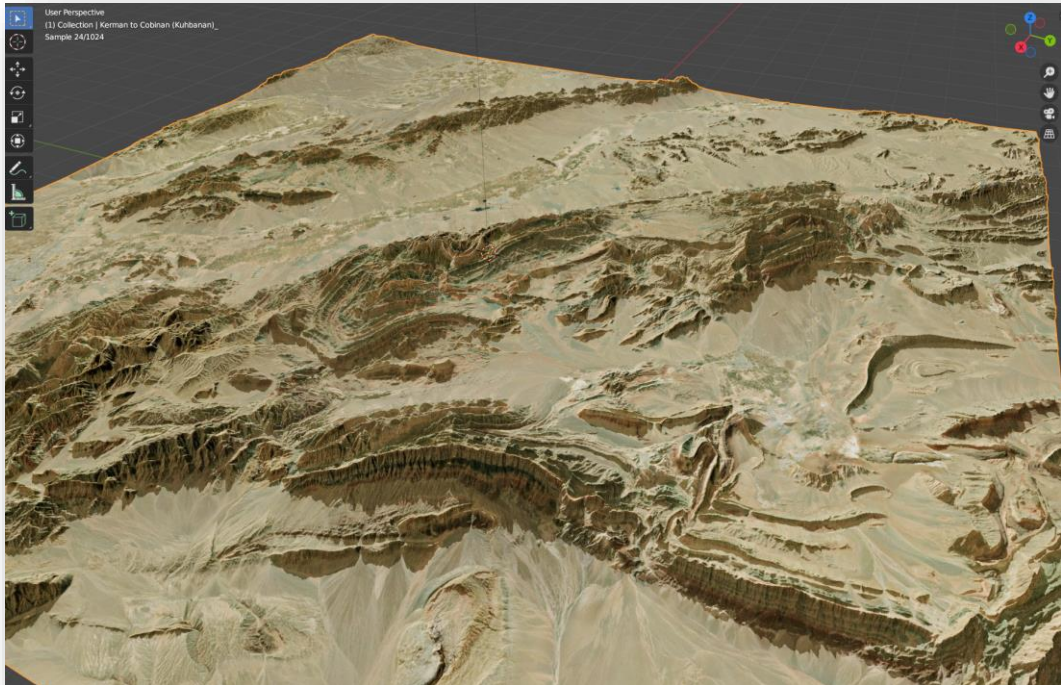
# SPATIAL ANALYSIS

## APPROACH

- Create a **bounding box for a partial route**
  - Kerman: (30.27305095, 57.0662499)
  - Cobinan (Kuhbanan): (31.4126295, 56.28006)
  - Tonocain (Tabas): (33.60953795, 56.9456578)
- **Download a DEM** (Digital Elevation Model) as GTIFF
  - COP30 (Copernicus Global DEM 30m)
- **Shade the generated 3D models** (*Blender, rayshader in R*)
  - Mapbox has been chosen as the source for satellite imagery (<https://www.mapbox.com/>)

# SPATIAL ANALYSIS

RESULT



# SPATIAL ANALYSIS

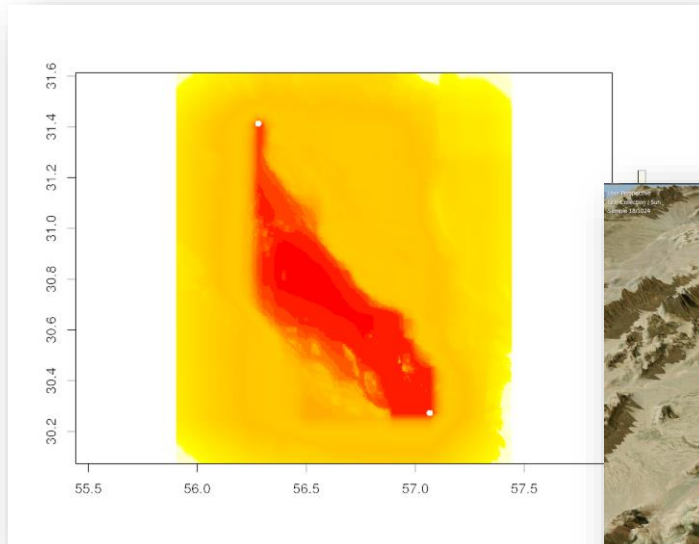
## COST PATH APPROACH

- As suggested by [10], "**Cost-Surface Analysis (CSA) and Least-Cost-Path Analysis (LCP) can be used to facilitate more nuanced interpretations of historical works** of travel writing and topographical literature".
- Initial data used to create the map is comprised of **two coordinates**:
  - The **origin** of the route's section and the **destination**.
  - **Mountains and valleys on the map can already be used as guides to guess a possible route that is easy to travel along.**
  - In order to back the readers' intuition, the **map can still be enhanced by highlighting the areas that are most easy** to travel through.

[10] Murrieta-Flores et al., 2017

# SPATIAL ANALYSIS

## RESULT



# SPATIAL ANALYSIS

## Cobinam to Tonocain

The Travels of Marco Polo

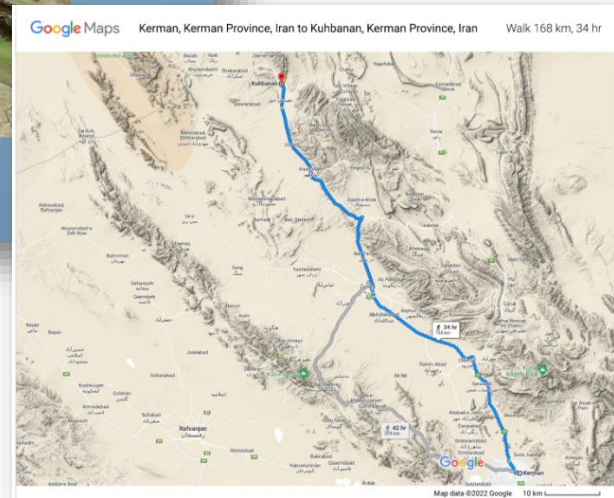


OF A CERTAIN DESERT THAT CONTINUES FOR EIGHT DAYS' JOURNEY.

When you depart from this City of Cobinan, you find yourself again in a Desert of surpassing aridity, which lasts for some eight days; here are neither fruits nor trees to be seen, and what water there is is bitter and bad, so that you have to carry both food and water. The cattle must needs drink the bad water, will they nill they, because of their great thirst. At the end of those eight days you arrive at a Province which is called TONOCAIN.




# SPATIAL ANALYSIS



# PIPELINE FOR TRAVEL LITERATURE

## RESULT

- NLP part must be improved further 
  - Using a pos-tagger, parser, plus VerbNet and FrameNet (VerbNet+FrameNet know about what structure it should be) we should be able to raise precision
- Generated Maps compared to Google Maps
  - While Google Maps provides a good overview of the route and the terrain's topography, the **renderings provide a more photo-realistic and immersive representation** of the route.
- Paths and Cost Corridors
  - Least Cost Path and **cost corridors show an optimal way to travel** but not necessarily a comfortable one.
  - Readers can **align their expectations with the descriptions** in the book

# FUTURE DIRECTION

## RESEARCH QUESTIONS 2.0

- Fact checking using cost corridors
  - Is the described **travel time realistic/plausible**?
  - Are the **thoughts of the reader** while receiving the text in **align with the actual situation**? (Psychology)
  - **Forensics applications** or Fact checking
- If they must have crossed landmarks like mountains, why are they not described in the text?
  - Maybe it was mentioned earlier: **Further linking of text parts using geo information**
- Study if such **maps help to mentally take in the entirety of the journey**