**Prompt-Engine**

Der Prompt ist eine zentrale Schnittstelle zwischen Menschen und Maschine. Er muss nicht auf eine Aufgabezeile reduziert werden, sondern kann mehrseitige Texte, Bilder oder gesprochene Eingaben enthalten. Schlüsselwörter ermöglichen semantische Bedeutung und Zusammenhänge im Text zu erkennen. So kann eine konsistente und präzise Antwort erzielt werden (vgl. Bünnagel 2024: 177).

Ein Bild, das Text, Screenshot, Schrift, Reihe enthält.

KI-generierte Inhalte können fehlerhaft sein.

Abb. 5.8 Ein effektiver Prompt ([Caelen/Blete 2004: Chapter 4](https://learning.oreilly.com/library/view/developing-apps-with/9781098168094/ch04.html))

Prompts dienen zu der Wahrscheinlichkeit bestimmter Antworten zu steuern. Sie fungieren als Leitfaden, der das Modell zu einer bestimmten Antworttype lenkt. Obwohl es keine standardisierte Struktur für Prompts existiert, bietet die Kombination aus Rolle, Kontext und Aufgabe ein hilfreiches Orientierungsschema (Abbildung 5.8). Zuweisung einer Rolle ist eine Möglichkeit, ein LLM gezielt zu steuern. Der Kontext bildet dabei das wichtigste Element eines Prompts. Bei der Formulierung eines Eingabetexts für LLM sollte der Kontext möglich präzise beschrieben werden. Dies erhöht die Chance, eine relevante Antwort zu bekommen. Die Erstellung eines Prompts ist ein iterativer Prozess, der oft mehrere Versuche erfordert. Zudem muss die Aufgabe klar und spezifisch definiert werden, wobei die relevanten Informationen klar und eindeutig bereitgestellt werden müssen (vgl. [Caelen/Blete 2004: Chapter 4](https://learning.oreilly.com/library/view/developing-apps-with/9781098168094/ch04.html)). Aufgrund dessen wurden im Rahmen dieser Arbeit verschiedene Prompts erprob, um die Effektivität der Ergebnisse zu erhöhen. So wurden bspw. die unterschiedlichen Beispiele im Kontextteil definiert und mehrere Aufgaben integriert (u.a. RE und die Normalisierung von Datumsangaben). Diese Vorgehensweise führte jedoch zu einer deutlichen Belastung des Modells, wodurch sich die Bearbeitungszeit erheblich verlängerte und sie zwischen 8 und 11 Stunden lag. Darüber hinaus wurde bei der Steuerung des Grads an Zufälligkeit in den Modellantworten sog. Temperatur experimentiert und am besten passender Wert von 0,7 rausgesucht. Die Werte dieses Parameters liegen zwischen 0 und 1: Je höher der Wert ist (1,0), desto abwechslungsreicher und kreativer sind die Ergebnisse (vgl. [LangChain](https://python.langchain.com/docs/concepts/chat_models/#standard-parameters)).

**Evaluieren!!!**

Nach der RE wurden die Tripel für jeden Satz überprüft und ggf. manuell korrigiert. Da Wikidata als Zielrepositorium dienen sollte, war es erforderlich, bestimmte Relationen in umgekehrter Form zu modellieren. Dies betraf insbesondere asymmetrische Relationen wie die Beziehungen Vater/Mutter – Kind, Ehrpartner:innen. Bei der manuellen Korrektur traten einige Probleme auf. Während der Datenvorverarbeiten wurden die Texte in einzelne Sätze untergeteilt, wodurch das Problem mit der Kohärenz entstanden. Ein Hauptgrund dafür war, dass die Entity Linking nur später durchgeführt wurde und nicht parallel mit NER. Dies wurde als Fehler im Nachhinein herausgestellt: Während die Verwendung des Personalpronomens *ich* unproblematisch war, da es eindeutig zu der Autorin oder dem Autor zugeordnet werden konnte, führte ein Wechsel der Erzählperspektive von Ich- zu Er-/Sie-Form zu Mehrdeutigkeit. Auch beim Gebrauch der ersten Person Plural (*wir*) musste jeweils geprüft werden, auf wen sich die Referenz bezog, wenn dieser nicht aus dem Kontext erkennbar war. Darüber hinaus wurden die vorgegebenen Relationen bei der RE ignoriert. Stattdessen erfolgte die Extraktion entweder aus dem Text direkt oder durch Paraphrasierung. Dies führte zu den unterschiedlichen Formulierungen für dieselbe Relation. Ein weiteres Problem war nur eine erkannte oder fälschlicherweise duplizierte Relation wie *X und Y sind Kinder von Z*. Dieses Problem musste auch manuell behoben werden.

**FP (Tätigkeit) wurde übernommen**

## 5.4.3 Entity Linking

Wissensgraphen wie Wikidata speichern strukturierte Informationen über Personen, Orte oder Konzepte. Jede Entität besitzt definierte Eigenschaften und Relationen. Diese Daten erleichtern semantische Anwendungen, z.B. Suchmaschine, die nach einer bestimmten Person suchen können. Das Hauptproblem dabei ist die Mehrdeutigkeit von Entitäten (vgl. Sevgili et al. 2022: 528). Die Erkennung dieser Entitäten ist nicht trivial. Gleiche Entitäten können vielen Name haben (Synonyme, Abkürzungen, Sprachvarianten) und umgekehrt ein Name steht für mehrere Entitäten (z. B. Neustadt) (vgl. Ege/Paschke 2021: 103). Die automatische Zuordnung von Entitäten zu einem Graph erfolgt durch Entity Linking (EL) (vg. Hamann Hinkelmann/Hoppe/Humm 2025: 83). Dadurch wird Mehrdeutigkeit aufgelöst (vgl. Sevgili et al. 2022: 528).

EL bekannt auch als Named Entity Disambiguation (NED) oder Entity Disambiguation. Sie etablierten sich im Forschungsfeld der Information Retrieval, natürlicher Sprache, semantisches Web und von den Datenbanken. Grundsätzlich umfasst EL zwei Aufgaben: die Erkennung von Entitäten im Text und die Disambiguierung. Bei der Disambiguierung werden Entitäten mit Strukturen und Wissensbasen wie Wikidata verknüpft werden (vgl. Mulang et al. 2020: 1–2).

Da die Maschinen keine indirekten Bedeutungen erkennen können, müssen Relationen zwischen Informationen eindeutig sein. Normdaten wie [GND-Nummer](https://www.dnb.de/DE/Professionell/Standardisierung/GND/gnd_node.html) (Gemeinsame Normdaten) ermöglichen diese Eindeutigkeit. Wird eine Person über GND-Nummer identifiziert, können Inhalte automatisch mit bspw. [Wikipedia](https://www.wikipedia.de/)-Eintrag oder [Deutschen Biographie](https://www.deutsche-biographie.de/) verknüpft werden (vgl. Kamzelak 2018: 426-427). Normdatensätze sind meistens über persistente Identifikatoren referenzierbar. Im Verknüpfungsprozess wird zunächst durch Disambiguierung der passende Normdatensatz aus den wahrscheinlisten Kandidaten bestimmt. Danach erfolgt die Verknüpfung der Entitätszeichenkette (vgl. Menzel et al. Franke-Maier et al. 2021: 231). Im Rahmen dieser Arbeit wurde der Prozess der Disambiguierung mithilfe der Wikidata-API umgesetzt.

Für die weitere Arbeit wurde die korrigierte JSON-Datei mit den Tripeln in eine Excel-Tabelle konvertiert. Zunächst wurden unterschiedliche Formulierungen derselben Relation auf ein einheitliches Wikidata-Prädikat abgebildet. Danach erfolgte mithilfe Wikidata-API die automatische Zuordnung der gewonnenen Subjekte, Prädikate und Objekte zu den entsprechenden Q- und P-IDs. Informationen, die nicht gefunden wurden, blieben leer und wurden händig ergänzt. Die Entitäten, die nicht identifiziert werden konnten, wurden aus dem Hilfstabelle entfernt.

Für die Erstellung eines neuen Items für eine Person wurde ein Mindestkriterium definiert: Jedes neue Item muss den vollständigen Namen der Person sowie mindestens eine zusätzliche Information enthalten, wie bspw. geographisches Attribut oder eine familiäre Relation. Dies dient zugleich der Vermeidung von Redundanzen und der Sicherung der Datenqualität. In der Praxis wird aus Effizienzgründen oder unter Zeitdruck nicht selten darauf verzichtet, bereits bestehende Datensätze auf ihre Richtigkeit und Vollständigkeit hin zu prüfen. Stattdessen wird ein neuer Eintrag erzeugt, was wiederum zu Redundanzen sowie zu einer Verringerung der Datenkonsistenz und -qualität im Gesamtsystem führen kann (vgl. Kamzelak 2018: 426-427). So wurden einige verdoppelte Q-IDs entdeckt, die in Zukunft noch korrigiert werden sollten. Darüber hinaus wurde die Hilfstabelle für die spätere Visualisierung um geografische Koordinaten erweitert, indem jedem Objekttype *location* die entsprechenden Werte zugewiesen wurden.

**Weg zum Missionsort**

**Wie wurden Prädikate von Tabelle zum Wiki-Datenmodell zugeordnet (Verben vs. P-IDs)**

## Wikidata-Datenmodell

Wie bereits im Kapitel 4 gesagt wurde, ist Wikidata eine multilinguale Graph-Datenbank, in der sowohl der Inhalt als auch Struktur des Graphen bearbeitet werden können. Da die Daten in Wikidata als RDF-Tripeln gespeichert werden, lässt sich das Datenmodell von Wikidata auf einfacher Ebene gut mit RDF abbilden: mit Item (rdf:subject), Property (rdf:predicate) und Wert (rdf:object) (vgl. [RDF und Wikidata](https://www.wikidata.org/wiki/Wikidata:Relation_between_properties_in_RDF_and_in_Wikidata)). Im Rahmen dieser Arbeit wurde zwei Typen von Items verwendet: Person mit biographischer Datensätze und Weg zu Missionsorten.

Da sich die Informationen über die Personen nicht als homogen zeigten, wurde sämtliches verfügbares Material im Datenmodell aufgenommen.

|  |  |  |
| --- | --- | --- |
| Wikidata-Property | Beschreibung | Beispiel |
| rdfs:label (@de) | *Label* des Items (Deutsch, Literal) | Johann Christian Breutel (wd:Q21415674) |
| wdt:P31 | *Instanz* von (verlinkt) | Mensch (wd:Q5) |
| wdt:P19 | *Geburtsort* (verlinkt) | Weißenburg in Bayern (wd:Q44500) |
| wdt:P20 | *Sterbeort* (verlinkt) | Berthelsdorf (wd:Q502553) |
| wdt:P21 | *Geschlecht* (verlinkt) | männlich (wd:Q6581097) |
| wdt:P22 | *Vater* (verlinkt) | Johann Christian Breutel (wd:Q21415674) |
| wdt:P25 | *Mutter* (verlinkt) | Sophie Breutel (wd:Q125843987) |
| wdt:P26 | *Ehrpartner(in)* (verlinkt) | Sophie Breutel (wd:Q125843987) |
| wdt:P40 | *Kind(er)* (verlinkt/ Literal) | 3 |
| wdt:P106 | *Tätigkeit* (verlinkt) | Botaniker (wd:Q2374149) |
| wdt:P463 | *Mitglied von* (verlinkt) | Herrnhuter Brüdergemeine (wd:Q159318) |
| wdt:P551 | *Wohnsitz* (verlinkt) | Berthelsdorf (wd:Q502553) |
| wdt:P569 | *Geburtsdatum* (Literal), ISO: YYYY-MM-DD/ YYYY-MM/ YYYY | 1788-01-21 |
| wdt:P570 | *Sterbedatum* (Literal), ISO: YYYY-MM-DD/ YYYY-MM/ YYYY | 1875-02-18 |
| wdt:P793 | *Schlüsselereignis* (verlinkt) | Weg zum Missionsort (Südafrika) (wd:Q135874728)/ Konfirmation (wd:Q214802) |
| wdt:P937 | *Wirkungsort* (verlinkt) | Südafrika (wd:Q258) |
| wdt:P3373 | *Geschwister* (verlinkt) | Johann Christian Breutel (wd:Q21415674) |
| pq:P854 | *URL der Fundstelle* (URL) | URL |

|  |  |  |
| --- | --- | --- |
| Wikidata-Property | Beschreibung | Beispiel |
| rdfs:label (@de) | *Label* des Items (Deutsch, Literal) | Weg zum Missonsort (Südafrika) (Q135874728) |
| wdt:P31 | *Instanz* von (verlinkt) | Route (Q1322323) |
| wdt:P3712 | *angestrebtes Ziel* (verlinkt) | Südafrika (wd:Q258) |
| wdt:P3342 | *relevante Person* (verlinkt) | Johann Gottlieb Bonatz (1773-1827) (wd:Q123396429) |
| pq:P854 | *URL der Fundstelle* (URL) | URL |
| pq:1427 | *geografischer Startpunkt* (verlinkt) | Herrnhut (wd:Q165140) |
| pq:P1444 | *geografischer Endpunkt* (verlinkt) | Genadendal (wd:Q1501295) |
| pq:P2825 | *Zwischenpunkt einer Reise* (verlinkt) |  |
| pq:P585 | *Zeitpunkt/Stand* (Literal), ISO: YYYY-MM-DD/YYYY-MM/YYYY | 1822-06-08 |
| pq:P580 | *Startzeitpunkt* (Literal), ISO: YYYY-MM-DD/YYYY-MM/YYYY | 1822-06-08 |
| pq:P582 | *Endzeitpunkt* (Literal), ISO: YYYY-MM-DD/YYYY-MM/YYYY | 1822-06-08 |

Person

Weg zum Missionsort (Südafrika, Labrador und Westindien)

<https://www.wikidata.org/wiki/Wikidata:Relation_between_properties_in_RDF_and_in_Wikidata>

## Semantic Web Technologie

<https://www.wikidata.org/wiki/EntitySchema:E49>

(Lanquillon/Schacht 2023, S. 187)

## SPARQL

Laube-Artikel

## Visualisierung

Netzwerkanalyse biete neue Möglichkeit zur Visualisierung und Analyse von Sammlungswissen. Ihre zentralen Ziele sind die Identifikation wichtiger Verbindungen und Strukturen. (vgl. Werner 2020: 248–249).

Für geistes- und sozialwissenschaftliche Forschung ist die Verbindung zwischen Netzwerken und Visualisierung von großer Bedeutung. Die graphische Visualisierung von abstrakten oder komplexen Strukturen macht die Zusammenhänge, die sich auf rein sprachlichem Wege nur schwer oder unzureichend vermitteln lassen, greifbar. Digitale Werkzeuge bieten hier neue Möglichkeiten, große Datenbestände nicht nur zu verarbeiten, sondern auch visuell aussagekräftige Formen überführen können. Daraus ergibt sich einen Anknüpfungspunk für die Verwendung netzwerkbasierter Modelle. Für die biographische Forschung eröffnet der Netzwerkansatz die Perspektive, individuelle Lebensläufe nicht isoliert zu betrachten, sondern als Teil größer Beziehungsgeflechte (vgl. Dörpinghaus 2022).

## Ergebnisse

## Auswertung / Perspektive

**OCR**: Zur Verbesserung der Ergebnisse könnte mehrstufige Vorgehensweise: Mehrere OCR-Durchläufe mit variierenden Parametern von Schwellenwerten bspw. und ggf. deren spätere Konsolidierung. Auf diese Weise lassen sich Fehlertypen gegenseitig kompensieren. Darüber hinaus konnte auch adaptive Vorverarbeitung helfen. Man könnte die Bilder sauber/ verschmutz, hell/dunkel annotieren und pro Klasse geeignete Parameter wählen. Als Nachkorrektur kann ein LLM eingesetzt werden.

**XML**: mehr Attributen zu definieren, Versionierung zeigen, Personen und Körperschaft gleich mit Wikidata Normdaten kodieren, Datumsangaben zu normieren,

## Zusammenfassung

## Literaturverzeichnis

## Anhang