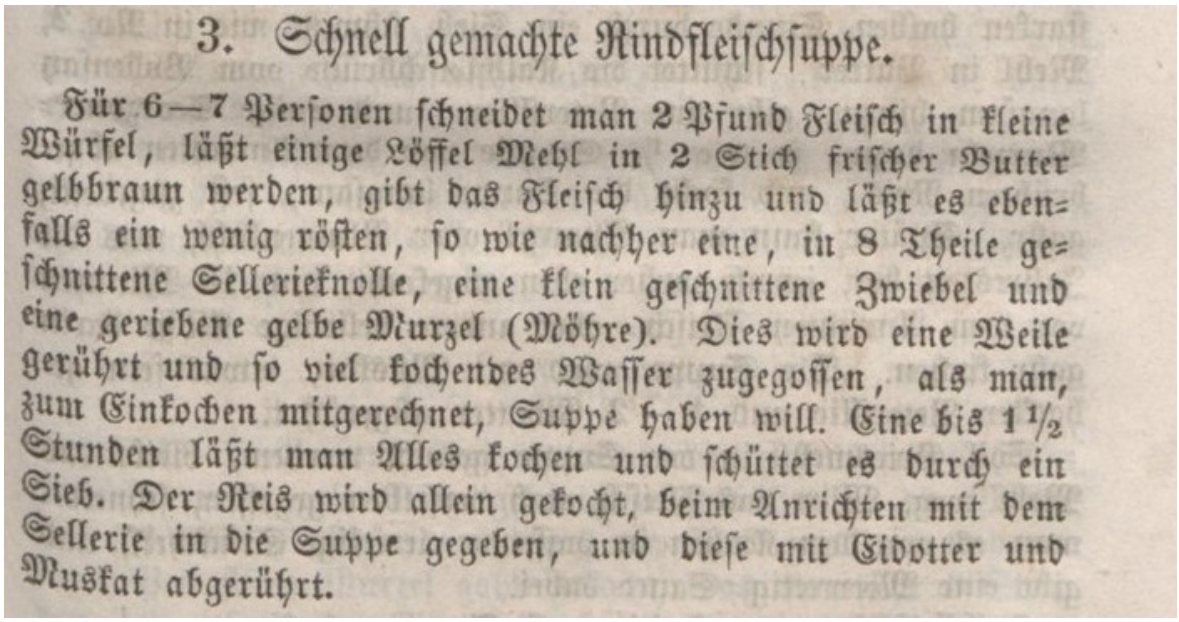
Problemstellung

- Beitrag zur Forschung
- Wie diese Arbeit zu lesen ist

Ziel dieser Arbeit ist die Erstellung einer digitalen Edition des Buches *Praktisches Kochbuch für die gewöhnliche und feinere Küche* (Davidis, 1849) (Quellen.html#DavidisKochbuch) welche für kulinarische Analysen genutzt werden kann. Eine Transkription des Kochbuches (http://www.deutschestextarchiv.de/book/view/davidis_kochbuch_1849?p=7) wurde bereits vom Deutschen Textarchiv angefertigt. Ein beispielhafter Auszug ist in Abb. 1 zu sehen. Abb. 1a zeigt den Scan eines Rezeptes, Abb. 1b die textuelle Form und Abb. 1c eine bereits von uns überarbeitete Version der TEI (http://www.tei-c.org/index.xml)-basierten Transkription.

Abb. 1: Transkription von Davidis's Kochbuch

Abb. 1a (Deutsches Textarchiv (A), 2008) (Quellen.html#DavidisKochbuchDTA)



3. Schnell gemachte Rindfleiſchſuppe.

Für 6—7 Perſonen ſchneidet man 2 Pfund Fleiſch in kleine Würfel, läßt einige Löffel Mehl in 2 Stich friſcher Butter gelbbraun werden, gibt das Fleiſch hinzu und läßt es ebenfalls ein wenig röſten, ſo wie nachher eine, in 8 Theile geſchnittene Sellerieknolle, eine klein geſchnittene Zwiebel und eine geriebene gelbe Murzel (Möhre). Dies wird eine Weile gerührt und ſo viel kochendes Waſſer zugegoſſen, als man, zum Einkochen mitgerechnet, Suppe haben will. Eine bis 1½ Stunden läßt man Alles kochen und ſchüttet es durch ein Sieb. Der Reis wird allein gekocht, beim Anrichten mit dem Sellerie in die Suppe gegeben, und dieſe mit Eidotter und Muskat abgerührt.

Abb. 1b (Deutsches Textarchiv (A), 2008) (Quellen.html#DavidisKochbuchDTA)

3. Schnell gemachte Rindfleiſchſuppe.

Für 6—7 Perſonen ſchneidet man 2 Pfund Fleiſch in kleine Würfel, läßt einige Löffel Mehl in 2 Stich friſcher Butter gelbbraun werden, gibt das Fleiſch hinzu und läßt es ebenfalls ein wenig röſten, ſo wie nachher eine, in 8 Theile geſchnittene Sellerieknolle, eine klein geſchnittene Zwiebel und eine geriebene gelbe Murzel (Möhre). Dies wird eine Weile gerührt und ſo viel kochendes Waſſer zugegoſſen, als man, zum Einkochen mitgerechnet, Suppe haben will. Eine bis 1½ Stunden läßt man Alles kochen und ſchüttet es durch ein Sieb. Der Reis wird allein gekocht, beim Anrichten mit dem Sellerie in die Suppe gegeben, und dieſe mit Eidotter und Muskat abgerührt.

Abb. 1c

```
<cue:recipe type="Suppen." rcp-id="B-3">
   <head>Schnell gemachte Rindfleischsuppe.</head>

   <p>Für 6–7 Personen schneidet man 2 Pfund
      Fleisch in kleine Würfel, läßt einige Löffel
      Mehl in 2 Stich frischer Butter gelbbraun
      werden, gibt das Fleisch hinzu und läßt es
      ebenfalls ein wenig rösten, so wie nachher
      eine, in 8 Theile geschnittene
      Sellerieknolle, eine klein geschnittene
      Zwiebel und eine geriebene gelbe Murzel
      (Möhre). Dies wird eine Weile gerührt und so
      viel kochendes Wasser zugegossen, als man,
      zum Einkochen mitgerechnet, Suppe haben will.
      Eine bis 1½ Stunden läßt man Alles kochen und
      schüttet es durch ein Sieb. Der Reis wird
      allein gekocht, beim Anrichten mit dem
      Sellerie in die Suppe gegeben, und diese mit
      Eidotter und Muskat abgerührt. </p>
</cue:recipe>
```

Für kulinarische Analysen sind insbesondere Zutatenlisten wünschenswert, welche bei den Rezepten nicht vorhanden sind. Für eine maschinenlesbare Aufarbeitung wird ein **Domänen-spezifisches Vokabular** benötigt. Idealerweise können aus den mit dem Vokabular ausgezeichneten Rezepten anschließend die Zutatenlisten automatisch extrahiert werden. Da die manuelle Auszeichnung zeitaufwendig und fehleranfällig ist, forschen wir des Weiteren im Bereich **Information Extraction** an Möglichkeiten zum automatischen Auszeichnen. Beides sind keine trivialen Probleme.

Domänen-spezifisches Vokabular

Bestehende Vokabulare ermöglichen keine kulinarische Analyse. TEI (http://www.tei-c.org/index.xml) hat kein spezifisches Vokabular um Rezepte auszuzeichnen. Schema.org/Recipe (https://schema.org/Recipe) ist ein Vokabular, welches das Ziel hat, Treffer von Suchmaschinen anzureichern. Es wird davon ausgegangen, dass jedes Rezept eine Zutatenliste hat. In dieser wird jede Zeile beispielsweise mit Microdata angereichert; z. B. *2 EL Zucker*. Für den Computer stellt dieser Tag-Inhalt jedoch nur einen String dar. Für eine kulinarische Analyse müsste die Mengenangabe (2), die Einheit (EL), sowie die Zutat (Zucker) explizit ausgezeichnet sein. Das Auszeichnen von Zutaten im Fließtext, was wir mangels bestehender Zutatenliste machen müssen, muss darüber hinaus einige Sonderfälle beachten. Dies verdeutlichen folgende zwei Beispielsätzen aus dem Kochbuch:

1. *„Der [Englische] Soja macht die Suppe gewürzreicher, kann jedoch gut wegbleiben, und statt Madeira kann man weißen Franzwein und etwas Rum nehmen.“* (Davidis, 1849, S. 33 f.) (Quellen.html#DavidisKochbuch)
2. *„Das Kalbfleisch wie in No. 1, nach der Personenzahl, doch etwas reichlicher genommen, da solches weniger Kraft gibt, als Rindfleisch.“* (Davidis, 1849, S. 30) (Quellen.html#DavidisKochbuch)

In Erstens ist Soja eine optionale Zutat. Je nachdem ob sie verwendet wird, hat das Gericht eine

unterschiedliche Geschmacksrichtung. In das Rezept gehören auch nicht Madeira, weißer Franzwein und Rum, sondern Madeira oder weißer Franzwein und Rum. Somit ist es nötig, bei der Auszeichnung zwischen **optionalen** und **alternativen Zutaten** zu unterscheiden. Des Weiteren sind alle drei Zutaten vage Begriffe. Rum gibt es zum Beispiel in vielen verschiedenen Preisklassen und mit unterschiedlichen Geschmacksausprägungen. Sofern vorhanden ist daher eine Präzisierung wünschenswert. Idealerweise wären **Zutaten bereits weltweit frei verfügbare und eindeutige Ressourcen mit abrufbaren Nährwertangaben**. Ein Mapping zwischen den verwendeten Zutaten pro Rezept und so einem Ressourcen-Bestand würde eine kulinarische Auswertung leicht und transparent machen. Leider gibt es unseren Wissens nach solche Ressourcen noch nicht.

In Zweitens ist Rindfleisch keine Zutat, sondern dient nur als Vergleich für eine ungefähre Mengenangabe. Wenn dieser Satz nur schnell gelesen wird, wird Rindfleisch hingegen leicht als Zutat eingeordnet (mir selber ist es erst beim dritten Mal lesen aufgefallen, dass es gar keine Zutat ist). Dies zeigt zum einen, dass beim Auszeichnen leicht Fehler gemacht werden können und zum anderen, dass nicht alle Zutaten als Zutaten des Rezepts auszuzeichnen sind.

Information Extraction in der Koch-Domäne

Information Extraction beschäftigt sich damit, nützliche Informationen aus unstrukturiertem Text zu extrahieren. Es wurde erstmals in (Feldman, Dagan, 1995) (Quellen.html#KDT) als eigene Disziplin erwähnt. Ein allgemeiner Überblick kann in (Hohto et al., 2005) (Quellen.html#TextMining) gefunden werden. Die Algorithmen versuchen meist bestehende Semi-Strukturen auszunutzen. In der Koch-Domäne haben wir drei unterschiedliche automatisierte Ansätze gefunden, auf die wir genauer eingehen werden:

- Regular Expression-based
- Conditional Random Field-based
- Dictionary- and rule-based

Nachdem die Zutaten von einem Programm extrahiert wurden, müssen die extrahierten Entities zum Auszeichnen nur noch entsprechend dem Domänen-spezifischen Vokabular zurückgeschrieben werden.

Unten stehend ist zur Veranschaulichung eine interaktive Grafik. Diese zeigt alle nötigen Schritte und Zwischenergebnisse von der Transformation des gedruckten Buchs bis hin zur ausgezeichneten digitalen Version als Webseite. Als erstes muss das gedruckte Buch transkribiert werden. Anschließend entwickeln wir eine Auszeichnungssprache, die eine kulinarische Analyse ermöglicht. Danach beschäftigen wir uns mit Programmen, welche die Rezepte automatisch auszeichnen. Abschließend werden die ausgezeichneten Rezepte in eine Webseite umgewandelt. Aufbauend auf dieser kann eine kulinarische Analyse durchgeführt werden, was nicht mehr Teil dieser Informatik-Arbeit ist. Wenn in dem jeweiligen Arbeitsschritt eine Transformation stattgefunden hat, ist rechts die Grundlage zu sehen und links, die neue, transformierte Version.

Beitrag zur Forschung

Der Beitrag dieser Arbeit zu dem aktuellen Forschungsstand lässt sich wie folgt zusammenfassen:

- Die Feststellung, dass bestehende Auszeichnungssprachen nicht für eine kulinarische

Analyse geeignet sind
- Entwicklung einer Auszeichnunggssprache für kulinarische Analysen
- Die Erkenntnis, dass Information Extraction mittels Conditional Random Fields für unsere Problemstellung nicht zielführend ist
- Erstellung eines dictionary- and rule-based Prototypen
- Dokumentation sowie Durchführung aller nötigen Arbeitsschritte, wie ein gedrucktes Buch digital zu einer Webseite aufgearbeitet werden kann

Wie diese Arbeit zu lesen ist

Diese Webseite ist das Ergebnis meiner Master-Arbeit mit dem Titel *Extracting recipe ingredients from cookbooks*.

Abb. 2a zeigt die chronologische Reihenfolge, in welche diese Arbeit zu lesen ist. Die Problemstellung wird am Anfang dieser Seite erläutert. Unter dem Menüpunkt Digitale Edition sind die weiteren Abschnitte zu finden. Der erste Abschnitt behandelt aufbauend auf der Version des Deutschen Textarchivs unsere Transkription des gedruckten Buchs von Frau Davidis. Im zweiten Abschnitt stellen wir bestehende Auszeichnungssprachen in der Koch-Domäne vor. Da diese nicht für eine kulinarische Analyse geeinigt sind, definieren wir für diesen Zweck im dritten Abschnitt die Auszeichnungssprache cueML. Der vierte Abschnitt beschäftigt sich mit der Information Extraction in der Koch-Domäne, mit dem Ziel später Rezepte automatisch mit cueML auszeichnen zu können. Darauf aufbauend werden im fünften Abschnitt zwei Prototypen zur Extraktion von Entities in der Koch-Domäne entwickelt. Die Evaluation dieser Prototypen ist Gegenstand des sechsten Abschnitts. Der siebte Abschnitt enthält einige Kommentare unsererseits zu dieser Arbeit. Der Menüpunkt Rezepte zeigt die von uns aufgearbeiteten und in HTML transofrmierten Rezepte Frau Davidis'. Der Menüpunkt Blog gibt Einblicke in unsere Arbeitsweise. Diese beiden Menüpunkte sind als bonus Material für den interessierten Leser zu verstehen.

Abb. 2b zeigt, wie der Aufbau der Arbeit inklusive der Unterabschnitte als klassiches Inhaltsverzeichnis aussieht.

Abb. 2: Aufbau der Webseite

Abb. 2a: Navigationsbar

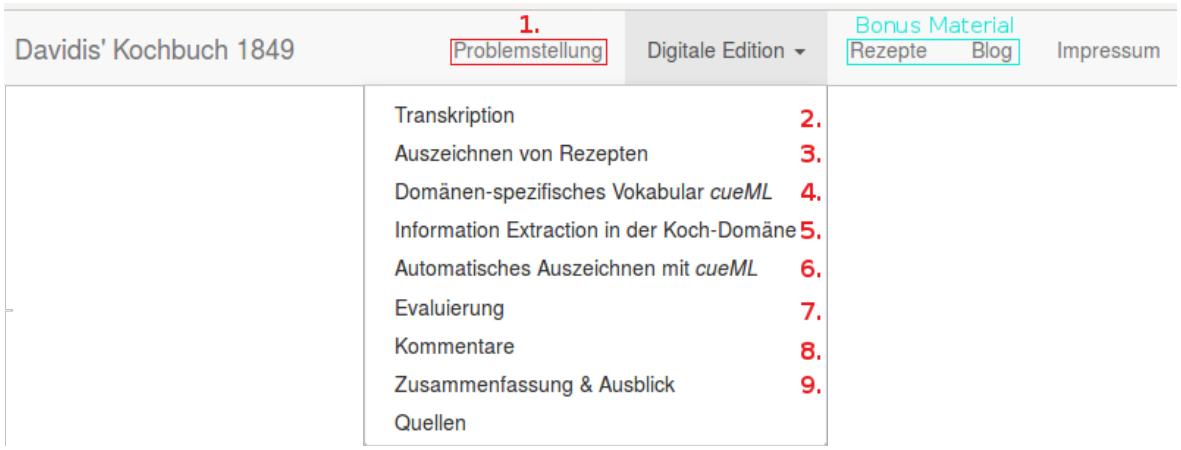


Abb. 2b: Klassisches Inhaltsverzeichnis