Universität Basel
Projektbericht
Wanja Gerber, Davide Miotto, Phoebe Misiuk, Gianna Coletti

Übung: Grand Tour gone Digital
FS 2024
Dozentin: Dr. Ina Serif

Projektbericht

Digitale Anreicherung und Geovisualisierung eines frühneuzeitlichen Reiseberichts

Charles Burney, The present state of music in France and Italy (1771)

Inhaltsverzeichnis

1. Einleitung

Das Projekt befasst sich damit, Teile des Reiseberichts *The present state of music in France and Italy* von Charles Burney aus dem Jahre 1771 aufzubereiten, anzureichern, und zu visualisieren.

Da der Text nur als Digitalisat vorhanden ist, wurde in einem ersten Schritt mit *Transkribus* ein Volltext erzeugt. Anschliessend wurde ein Named Entity Recognition-Modell für SpaCy trainiert und auf zwei Kapitel des Textes angewandt. Nach der Korrektur und Ergänzung der ausgezeichneten Entitäten konnte eine Geovisualisierung mithilfe von Nodegoat erstellt werden.

2. Transkribus

Die PDF-Datei des entsprechenden Kapitels, hier zum Kapitel “Venice”, konnte ohne Probleme auf *Transkribus* hochgeladen werden. Mithilfe des öffentlichen Modells “Transkribus Print M1” wurde dann die Texterkennung durchgeführt. Da das Modell auch auf englische Texte trainiert war, hatte es keine Probleme, das englische “f” mit einem “s” zu ersetzen. Nach der Texterkennung konnte der Text als Volltextdokument exportiert werden, um darauf aufbauend das Named Entity Recognition-Modell zu trainieren und schliesslich das gesamte Kapitel so zu annotieren. Für dieses Projekt wurden die Entitäten Date (<date>), Location (<place>), Person (<perName>), Instrument (<instrument>) und Music (<music>) festgelegt. Dies mit dem Ziel, ein Ereignis mit den annotierten Informationen auf einer Karte von Venedig zu lokalisieren.

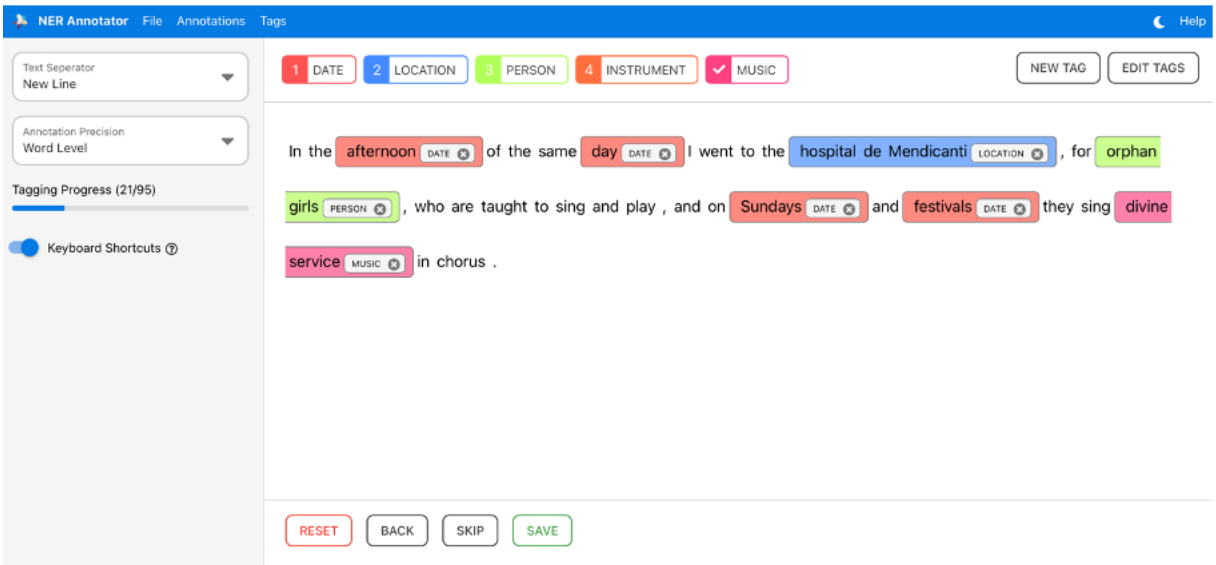
3. Named Entity Recognition mit SpaCy

3.1 Preprocessing

Für die folgenden Prozesse wurde ein Python-Skript verwendet, das in der Datei *01_preprocessing.ipynb* eingesehen werden kann. In der Transkription von *Transkribus* kommt dieser Verbindungsstrich häufig vor: ¬. Durch das Skript werden diese Zeichen entfernt und die dadurch getrennten Worte werden verbunden. Dann werden in den folgenden Schritten alle Seitenzahlen entfernt und der Text wird so formatiert, dass nach jedem Satzende eine neue Zeile beginnt. Dieser wird schliesslich als neue Datei abgespeichert. Anschliessend werden Zeichen, Sätze und Tokens im Datensatz gezählt, sodass später eine Aufteilung der Daten in Trainings- und Validierungsdaten stattfinden kann.

3.2 Modell trainieren

Nun ist der Text vorbereitet und wird im NER Annotator ausgezeichnet.



Die Annotations können als json-Datei exportiert werden. Diese werden nun in Trainings- und Validierungsdaten aufgeteilt (80/20). Der Code ist hier zu finden: 02_trainData-evaluationData.*ipynb*. Diese Dateien werden dann in das spacy-Format umgewandelt. Bei diesem Schritt sind anfangs einige Probleme entstanden. Das Skript konnte die Datei nicht umwandeln, da es in den Daten “null”-Werte gab. Nachträglich konnten diese darauf zurückgeführt, dass der Text im Preprocessing nicht ausreichend bereinigt und formatiert wurde, was beim Output des NER Annotator zur Entstehung der “null”-Werte führte. Schliesslich kann mithilfe der ausgezeichneten Daten das Training in der Kommandozeile durchgeführt werden.

```
(base) wanjagerber@MBP-von-Wanja downloads % python -m spacy train config.cfg --output ./output --paths.train /Users/wanja-gerber/Desktop/finalNER/trainData.spacy --paths.dev /Users/wanjagerber/Desktop/finalNER/valuationData.spacy
```

Der Output hat folgende Form:

```
✔ Created output directory: output
ℹ Saving to output directory: output
ℹ Using CPU

=========================== Initializing pipeline ===========================
✔ Initialized pipeline

============================= Training pipeline =============================
ℹ Pipeline: ['tok2vec', 'ner']
ℹ Initial learn rate: 0.001
E    #       LOSS TOK2VEC  LOSS NER  ENTS_F  ENTS_P  ENTS_R  SCORE
---  ------  ------------  --------  ------  ------  ------  ------
  0       0          0.00     15.64    0.00    0.00    0.00    0.00
  6     200        158.20   2506.56   74.96   71.02   79.37    0.75
 15     400         48.50    614.28   92.45   91.59   93.33    0.92
 25     600         29.84    145.35   92.96   91.67   94.29    0.93
 39     800         23.01     76.11   91.19   88.86   93.65    0.91
 55    1000         28.53     68.38   91.36   88.89   93.97    0.91
 76    1200         28.90     49.65   92.86   90.88   94.92    0.93
101    1400         99.82    114.13   91.69   92.28   91.11    0.92
132    1600         36.08     33.55   90.85   90.28   91.43    0.91
171    1800          9.92      8.11   91.71   90.43   93.02    0.92
217    2000          0.00      0.00   91.71   90.43   93.02    0.92
275    2200          0.00      0.00   91.48   90.91   92.06    0.91
✔ Saved pipeline to output directory
output/model-last
```

3.3 Evaluation

Mithilfe der folgenden Eingabe wird das Modell evaluiert:

```
(base) wanjagerber@MBP-von-Wanja downloads % python -m spacy package /users/wanjagerber/downloads/output/model-best /users/wanjagerber/downloads/package --build wheel
```

Unser Modell schneidet so ab:

```
================================== Results ==================================

TOK      100.00
NER P    90.73
NER R    93.35
NER F    92.02
SPEED    3668


============================== NER (per type) ==============================

                P       R       F
PERSON      92.13   92.86   92.49
LOCATION    91.14   96.00   93.51
MUSIC       96.97   86.49   91.43
DATE        92.86  100.00   96.30
INSTRUMENT  78.57   97.78   87.13
```

Die Evaluation zeigt, dass das Modell einen F-Score von 0.92 aufweist, also ein gutes Resultat.

3.4 Anwendung

Nach der positiven Evaluation kann das Modell angewandt werden. In der Datei 03_*Anwendung.ipynb* ist der entsprechende Code für die Anwendung auf ein Kapitel aus Bourneys Reisebericht einzusehen. Dabei wird auch hier im ersten Schritt der Text bereinigt: Entfernung der Bindestriche und Zusammenführung der Wörter, sowie Aufteilung in Absätze. Anschliessend muss die Text-Datei in eine xml-Datei umgewandelt werden, sodass diese wiederum als Grundlage für die NER dienen kann. Nach der erfolgreichen Anwendung können die Auszeichnungen in einer neuen xml-Datei in den Text eingefügt werden.

4. Anreicherung und Korrektur der TEI-XML-Datei

Zur Funktionalität für eine Visualisierung müssen örtlich und zeitlich zusammengehörige Entitäten in Events zusammengefasst werden. Nach Rücksprache mit Frau Serif war klar, dass wir die Events von Hand nachtragen mussten. Da der Textabschnitt zu Venedig mit 54 Seiten sehr umfangreich ist, beschränkten wir uns hier auf die ersten 17 Events. Zum Annotieren von Events wird am Anfang und Ende des jeweiligen Events im Text der Tag <event> mit der jeweiligen Nummer als id-Attribut <event id="event1"> platziert. So kann man Events spezifisch auf Xpather suchen und alle annotierten Elemente innerhalb dieses Events nach Entität geordnet auf die CSV-Datei exportieren. Das id-Attribut dient auch zur Zuordnung innerhalb der CSV-Tabelle, die mithilfe eines Python-scripts (siehe 04_export.ipynb) erstellt wird. Korrekturen der ausgezeichneten Entitäten wurden erst nach dem Export als CSV-Datei vorgenommen.

5. Export und Visualisierung

5.1. Export

Um die ausgezeichneten Elemente der XML-Datei bei Nodegoat importieren zu können, müssen diese im csv-Format vorliegen. Zunächst war Xpather zur Extraktion der Daten angedacht, die Copy-Paste-Methode erwies sich jedoch als nicht zielführend. Daher verwendeten wird das Python-Skript *Export-csv.ipynb*.

5.2. Koordinaten

Um die von Burney beschriebenen Orte auf einer Karte darstellen zu könnten, mussten sie vor dem Exportieren der CSV-Datei auf die Plattform Nodegoat mit Koordinaten versehen werden, was leider nur durch händische Arbeit möglich war. In fast allen Fällen konnte der im Text genannte Ort mittels moderner Kartendienste wie Google Maps lokalisiert werden, teils waren auch kurze Recherchen von Nöten, wenn Burney abgekürzte Ortsbezeichnung verwendete. Da es sich bei den Orten, an denen musiziert wurde, meist um Kirchen handelte, die es heute noch gibt, stellte die Lokalisation dieser keine allzu grosse Herausforderung dar. Google Maps ermöglichte es, Koordinaten im Format (*longitude* / *latitude*) von den Ort zu bestimmen, um die Ortsangaben in er CVS-Datei zu vervollständigen.

5.3. Schritte zur Visualisierung auf Nodegoat:

5.3.1. Objekt erstellen

Ein Objekt namens Event wurde erstellt. Darunter 5 Sub-Objekte:

- "People" - Beschreibt welche Personen beim Event anwesend waren.
- "Place" - Gibt der Name des Platzes, wo der Event stattfand an, sowie auch seine Koordinate (wenn möglich).
- "Date" - Gibt Charles Burneys Beschreibung und auch das Datum in tt/mm/jjjj Format an.
- "Instrument" - Beschreibt welche Instrumente am Event gespielt wurden.
- "Music" - Beschreibt der Typ von Musik, der gespielt wurde.

5.3.2. Projekt erstellen

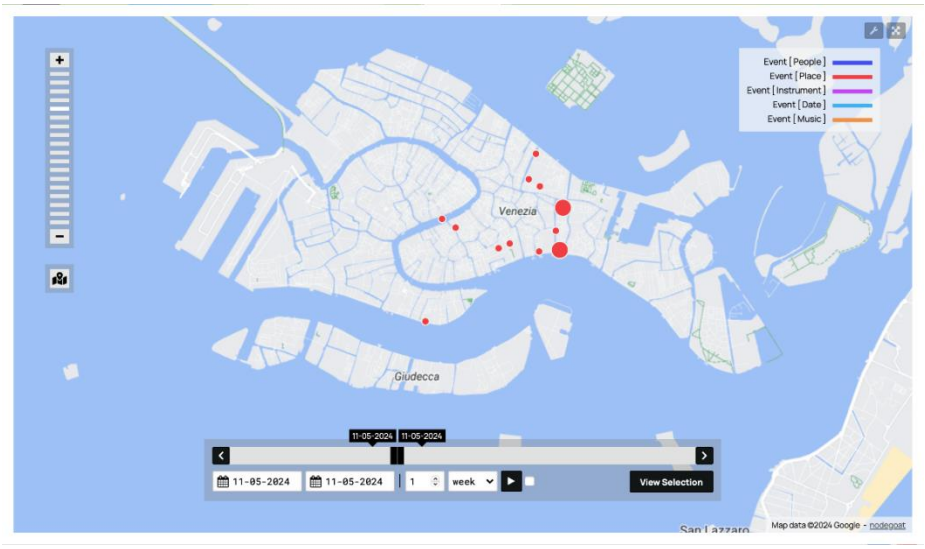
Das Projekt wurde "The present state of music in France and Italy - Charles Burney" genannt. Das Objekt "Event" wurde zum Projekt beigefügt.

5.3.3. CSV-Datei importieren

Erst wurde die CSV-Datei auf Nodegoat hochgeladen. Daraufhin wurde ein Import Template erstellt, indem die Spalten in der CSV-Datei mit dem Objekt und Sub-Objekte verknüpft wurden.

5.3.4. Visualisieren

Die verschiedenen Events wurden mit Nodegoat geografische visualisiert.



Erster Versuch einer Visualisierung der Events in Venedig.

5.3.5. Filter erstellen

... TBC

6. Probleme, Fragestellungen, Ausblick

6.1. Quellensuche

Diese Arbeit ist primär durch das Interesse am Bericht *The Present State of Music in France and Italy* von Charles Burney entstanden und darauf aufbauend entwickelt worden. Im Bericht beschreibt Burney Momente, in denen er während einer Reise durch Frankreich und Italien erlebt hat, wie Musik gespielt wurde. Der Bericht ist nach den Städten gegliedert, die er besucht hat. Die Quelle bot ideale Voraussetzungen zum Erlernen und Anwenden von geographischen Visualisierungen, da im Bericht oft der Ort in Verbindung mit einem jeweiligen Ereignis, an dem Musik gespielt wird, anschaulich erwähnt wird.

6.2. Aufbereitung der Quelle

Zunächst konzentrierten wir uns in allen Arbeitsschritten auf Burneys Berichte während seines Aufenthalts in Venedig, da sein gesamter Reisebericht zu umfangreich für diese Gruppenarbeit gewesen wäre.

Alle 54 Seiten von Burney Schilderungen zu Venedig konnten problemlos transkribiert werden. Als Entitäten, die später mittels Named Entity Recognition ausgezeichnet werden sollen, haben wir *Daten, Personen, Orte, Instrumente* und *Musikstücke* festgelegt. Zudem haben wir Möglichkeiten zum Export der Daten aus dem XML-Dokument besprochen, zunächst schien uns xpather eine gute Möglichkeit zu sein. Ebenfalls wurde diskutiert, welche Programme zur Visualisierung sinnvoll sein könnten. Unter anderem haben wir, auch nach Rücksprache mit Frau Serif, den Geobrowser in Betracht gezogen, welcher jedoch abgesehen von Ort-Zeit-Relationen keine anderen Darstellungen ermöglicht. Storymaps Knightlab wäre ebenfalls eine Option gewesen, erfordert jedoch sehr viel händische Arbeit, weshalb wir uns trotz langer Einarbeitungszeit für Nodegoat entschieden haben.

Ein Problem, was in verschiedenen Arbeitsschritten wiederkehrend auftrat, war die fehlerhafte Auszeichnung von Entitäten durch das NER-Modell. Nebst offensichtlichen Fehlauszeichnungen wurden auch Entitäten ausgezeichnet, die nicht für die Visualisierung benötigt wurden, wie beispielsweise “me” als Person. Dies und Wiederholungen von Entitäten innerhalb eines Events mussten von Hand in der CSV-Datei bereinigt werden. Zudem wurden auch Entitäten ausgezeichnet, die nicht direkt mit dem Event in Verbindung standen, sondern vom Autor als Referenz erwähnt wurden. So mussten auch diese händisch korrigiert werden, da kein automatisierter Lösungansatz dafür gefunden wurde.

Die Übersetzung von XML-Datei in eine CSV-Datei bot ein weiteres Problem. Die funktionen von Objekt und SUb-Objekt mussten auf Nodegoat abgestimmt sein, um die folgende Visualisierung zu ermöglichen. Dieses Problem wurde schliesslich mithilfe eines Python Scripts gelöst (siehe 04_export.ipynb).

6.3. Visualisierung mit Nodegoat

Als eine zentrale Schwierigkeit bei Nodegoat erwies sich die Animation der Zeitleiste, denn alle Events tauchen, obwohl sie als an unterschiedlichen Daten stattfindend kennzeichnet wurden, gleichzeitig auf der Karte auf.

6.4. Ausblick

Das Trainieren der Named Entity Recognition auf den Text von Burney erlaubt es, weitere Städte zu annotieren und schliesslich zu visualisieren. Somit kann man anhand dieser Visualisierung quantitative Fragestellungen zur Verbreitung von gewissen Instrumenten oder zur Art und Häufigkeit von gespielter Musik durchführen. Schon innerhalb von Venedig sind Vergleiche möglich, beispielsweise in welchen Quartieren am meisten musiziert wird. Weitere Fragestellungen könnten ebenfalls mit zusätzlichem ausgezeichnetem Material behandelt werden:

- Mit welchen Musikinstrumenten wurde Ende des 18. Jh. in [einer spezifischen Ortschaft bzw. zwei Ortschaften zum Vergleich oder grundsätzlich in Italien und/oder Südfrankreich] musiziert? In welchem Aufführungskontext wurden welche Musikinstrumente verwendet?
- Wie lassen sich Musizierende im 18. Jh. charakterisieren? Welche Rückschlüsse auf die Art und Weise der musikalischen Bildung im 18. Jh. in Italien und/oder Südfrakreich sind möglich?
- An welchen Örtlichkeiten wurde im Italien / Südfrankreich des 18. Jh. musiziert? Welche Orte wurden für welche Art des Musizierens genutzt?