Premodern data for NLP

Panel A2 - Basel
21.10.21, 15h00 - 15h45
Panel suggested by Jonas Widmer and Christa Schneider
Reporter: Gabi Wuethrich

[Deutsch, see English translation below]

Präsentation:
Die Leitfrage, die die beiden Panelverantwortlichen beschäftigt, ist das Verhältnis von Aufwand und Ertrag bei digitaler Aufbereitung prämoderner Daten. Konkret geht es um das Projekt zu den Berner Turmbüchern, in dem sie beide involviert sind.

Bei den Turmbüchern handelt es sich um Protokolle von Kriminalprozessen, die zwischen 1547 und 1798 mehr oder weniger chronologisch erhalten, aber z.T. unter Folter entstanden sind. Insgesamt wurden ca. 300'000 in Kurrentschrift geschriebene Seiten mit Transkribus erfasst, die nun aufwändig korrigiert werden. Die verschiedenen Schreiber können in den meisten Fällen identifiziert werden, da sie gewählt wurden. Zum Teil ist auch ihr Hintergrund bekannt.
Christa Schneider interessiert sich als historische Soziolinguistin für Fragen zu den verwendeten sprachlichen Mitteln, Dialektwörtern etc. In der Schweiz gab es keine sprachliche Normierung bis weit ins 18. Jahrhundert. Die Turmbücher können so zum Verständnis beitragen, wie sich die Berner Sprache zuvor entwickelt hat, und zum sprachlichen Alltag im Allgemeinen: Welche Schriftlichkeit wird verwendet, wie passen sich die Schreiber an?

Um die Texte auf diese Fragestellungen hin untersuchen zu können, muss ihre Struktur zunächst entsprechend erfasst werden. Dies soll idealerweise mit Hilfe von NLP geschehen. Jonas Widmer zeigt ein einfaches Beispiel unter Verwendung der Spacy Library De_core_news (https://spacy.io/models/de). Entsprechende Libraries existieren bislang zwar für moderne Sprachen, aber nicht für prämoderne - was eine Herausforderung für die Umsetzung darstellt.

Die wichtigsten Argumente in der Diskussion:
Grundsätzlich sind die automatisiert transkribierten Texte noch sehr unstrukturiert und uneinheitlich, was auch dem Zeitdruck des Schreibers geschuldet sein mochte. Um entsprechende NLP-NER-Tags zu trainieren, braucht es eine grosse Menge an Trainingsdaten, die händisch eingegeben werden müssen. Dabei stellt sich wie erwähnt die Leitfrage des Panels: Die für NLP benötigte Trainingsdatenmenge ist so gross, dass sich die Panelverantwortlichen unsicher sind, ob NLP überhaupt zielführend ist. So meint Christa, dass man bereits mit den Transkriptionen enorm viel anfangen könne. Es stelle sich auch die Frage, wer eine solche NLP-Library weiternutzen würde, weil die Anzahl Quellen für eine Weiternutzung begrenzt sei. Zudem seien die Trainings sehr rechenleistungsaufwändig.

Vorgeschlagene Lösungen:
Eine Idee, um den Aufwand zu verringern, ist, den Leipziger Kodex als bereits trainiertes Beispiel einzubinden und auf Berner Verhältnisse anzupassen. Zudem muss eine Balance zwischen dem mit Ungenauigkeiten verbundenen Informationsverlust und dem Aufwand für eine genaue Analyse gefunden werden. Ein weiteres Anliegen ist die generelle Gefahr maschineller Methoden, Bias in den Daten zu fördern. Die Panelinitator*innen hofften dabei auf die Erfahrungsberichte anderer Teilnehmenden.

Ein Vorschlag ist, die Basler Avis-Blätter als Trainingsbasis zu verwenden. Allerdings ist derzeit noch unklar, auf welcher Plattform diese veröffentlicht werden. Das "beeindruckende" Projekt erinnert einen Teilnehmenden an die Grundlagenforschung zu Transkribus. Bibliotheken seien auf diese angewiesen. Bei solch “riesigen” Projekten fielen auch immer Byproducts wie Tagger an, die weiterverwendet werden könnten. Für die Weiterverwendung des entsprechenden Tagger stellt sich die Frage, wie er zu veröffentlichen wäre.

Eine Möglichkeit für eine Weiterverwendung der Tagger wären e-rara und e-manuscripta, bei denen die Texterkennung nach wie vor eine Baustelle sei. Derzeit basiert die e-rara-Texterkennung auf einem kommerziellen Produkt, während bei e-manuscripta gerade ein Versuch mit Transkribus läuft. Grundsätzlich muss der Bedarf nach einer Texterkennung deutlich kommuniziert werden. Auch seien digitale Nutzer bei der Governance der e-manuscripta eher untervertreten. Ein Mehrwert der NLP-Methodik wäre im Idealfall z.B. die Personensuche in Bibliotheks- und Archivsystemen.

Mit Blick auf den von Jonas kritisch eingeschätzten Trade-off zwischen aufwändiger Präzision und Informationsverlust bleibt anzumerken, dass selbst mit weniger Präzision über NLP zusätzliche Informationen gefunden werden können, die zuvor nicht auffindbar waren. Eventuell könnten kostenpflichtige Dienstleister im Ausland einen Teil der Codierung übernehmen, wobei der "Datenkolonialismus" grundsätzlich problematisch sei.

Christa steht dem Vorschlag zu den kostenpflichtigen Anbietern kritisch gegenüber, weil die Textinhalte z.T. sehr sensibel und psychologisch schwierig zu lesen seien, auch wenn Personendaten kein Problem mehr darstellten. Es brauche ein gewisses Vorwissen, um mit den Texten klarzukommen, das man z.B. Studierenden besser vermitteln könne.
Zum Schluss gibt es mehr oder weniger ernsthafte Überlegungen, wie man die Codierung allenfalls über Gamifizierung spannender gestalten könnte.

Grundsätzlich müssen bei solch grossen Projekten stets die Vor- und Nachteile digitaler Erschliessungsmöglichkeiten gegeneinander abgewogen werden:
Vorteile sind:
- Die Automatisierung manuell nicht realisierbarer Projekte/Aufgaben
- Die Erstellung von Hilfsmitteln für die Forschung
- Digital aufbereitete Daten bleiben verfügbar und bieten einen Mehrwert für weiterführende Arbeiten

Als Nachteile gelten:
- Bei der Arbeit mit Texten ist ein linguistischer Informationsverlust durch die Normalisierung möglich

- Entsprechendes Know How, Infrastruktur und die Ressourcen müssen vorhanden sein
- Bei maschinellen Methoden stellen sich immer ethische Fragen: Datenbias, fehlende Kontextualisierung etc.

[EN, see DE original version above]

Discussion leader and core topic presentation:
The main concern of the two panel members that suggested this topic is the trade-off between effort and return of digitally processing pre-modern data. They specifically situate their questions in the context of the Bernese Tower Books, an edition project in which they are both involved.

The Tower Books are records of criminal trials that have been preserved more or less chronologically between 1547 and 1798, some of which were written under torture. A total of about 300,000 pages in Kurrent script have been transcribed with Transkribus, which are now being corrected with the utmost effort. The various scribes can be identified in most cases because they were elected. In some cases their background is also known.

Christa Schneider, as a historical sociolinguist, is interested in questions concerning the linguistic devices used, dialect words, etc. In Switzerland, there was no linguistic standardization until well into the 18th century. The Tower Books can thus contribute to the understanding of how the Bernese language developed before that, and to everyday linguistic life in general: What kind of writing is used, how do scribes adapt?

In order to be able to examine the texts with regard to these questions, their structure must first be recorded accordingly. Ideally, this should be done with the help of NLP. Jonas Widmer shows a simple example using the Spacy Library De_core_news (https://spacy.io/models/de). Corresponding libraries, at the moment, only exist for modern languages, but not for pre-modern ones - challenging their implementation in the project.

Main arguments in the discussion:
In general, the automatically transcribed texts are still very unstructured and inconsistent, which might also be due to the time pressure of the scribes. In order to train the respective NLP-NER tags, a large amount of training data is needed, which has to be entered manually. As mentioned, this raises the leading question of the panel: The amount of training data needed for NLP is so large that the panel members are unsure whether NLP is effective at all. Christa thinks that they can already do a lot with the transcriptions. There is also the question of who would continue to use such an NLP library, because the number of sources for further use is limited. In addition, the training needs intensive computational power.

Proposed solutions:
One idea to reduce the effort is to include the Leipzig Code as an already trained example, and to adapt it to Bernese conditions. In addition, a balance between the loss of information associated with inaccuracies, and the effort required for accurate analysis, must be found. Another concern is the general danger that machine learning methods promote bias in the data. In this regard, the panel initiators counted on the experience and testimonials of the participants.

One suggestion is to use the Basel Avis Blaetter as a training base. However, it is currently unclear on which platform these will be published. The "impressive" project reminds one participant of the basic research on Transkribus. Libraries depend on that basic research, he said. In such "huge" projects, byproducts such as taggers are produced that could be reused. How to publish such taggers for reuse, remains an open question, however.

One possibility for further use of the taggers would be records in e-rara and e-manuscripta, where text recognition is still an open issue. Currently, e-rara text recognition is based on a commercial product, while e-manuscripta is running a trial with Transkribus. In general, the need for text recognition needs to be clearly communicated. Also, digital users tend to be underrepresented in the governance of e-manuscripta. Ideally, an added value of NLP methodology would be, for example, people searches in library and archive systems.

With regard to Jonas's critical assessment of the trade-off between elaborate precision and loss of information, it remains to be noted that even with less precision, additional information can be found via NLP that was previously not searchable. Possibly, fee-based service providers abroad could take over some of the coding, although "data colonialism" is fundamentally problematic.

Christa is critical of the proposal regarding fee-based providers because the content content is sensitive and psychologically difficult to read, even if personal data is no longer a problem. It needs a certain basic knowledge of the context to cope with the texts, which is easier to convey to students for instance.

Finally, there are more or less serious considerations about how coding could be made more exciting via gamification.

Basically, with such large projects, the advantages and disadvantages of digital indexing possibilities must always be weighed against each other:

Advantages are:
• The automation of manually unfeasible projects/tasks.
• The creation of research tools.
• Digitally prepared data remain available and provide added value for further work
Disadvantages include:
• When working with texts, a linguistic loss of information due to normalization is possible
• Appropriate know-how, infrastructure, and the resources must be available
• Ethical issues always arise with machine learning methods: data bias, lack of contextualization, etc.