

Eine Fallstudie zur Annotation von Vagheit in Werken Dimitrie Cantemirs

Vertan, Cristina

cristina.vertan@uni-hamburg.de
Universität Hamburg, Deutschland

von Hahn, Walther

vhahn@informatik.uni-hamburg.de
Universität Hamburg, Deutschland

Das Korpus

Das ausgewählte Korpus besteht aus zwei Hauptwerken Dimitrie Cantemirs, eines Universalgelehrten des 17. Jahrhunderts und Mitglied der „Kurfürstlich – Brandenburgischen Societät der Wissenschaften“. Die zwei Werke wurden ursprünglich auf Lateinisch verfasst, die Originale sind aber verloren aber Kopien davon wurden im späten 20. Jh. wiederentdeckt. Im Umlauf waren lange Zeit nur Übersetzungen ins Englische, Deutsche (Cantemir 1771), (Cantemir 1745), und Französische, die mindestens bis Mitte des 19. Jh. Referenzwerke für die Geschichte des osmanischen Reichs und der historischen Provinz Moldawien waren. Durch seinen langen Aufenthalt in Istanbul, hatte Cantemir Zugang zu vielen Quellen die er zitiert. Daneben zitiert er auch Sagen und Legenden und versucht immer durch geschickte sprachliche Redewendungen zu vermitteln, was seiner Meinung nach historisch gesichert ist. Daher ist das Korpus besonders illustrativ für das Problem der Vagheitsannotation.

Erste Schritte zur Vagheitsannotation

Das Projekt HerCore versucht durch gezielte Annotation von Vagheit drei geisteswissenschaftliche Fragestellungen in Bezug auf die Cantemir-Forschung zu lösen:

- Der erstmalige Vergleich aller historischen Übersetzungen, da seit geraumer Zeit die Vermutung formuliert wurde, dass diese relativ stark von den Originalen abweichen.
- Die Untersuchung der Zuverlässigkeit von Äußerungen Cantemirs. Hierbei werden vor allem Quellen von turkologischen Fachwissenschaftlern einbezogen.
- Die Konsistenz von Cantemir über dieselben Personen und Ereignisse in den zwei Werken.

Die Annotation von Vagheit wird auf drei Ebenen untersucht:

1. Linguistisch,
2. In Metadaten und Editorik,
3. im Fachwissen.

- Für die Linguistische Ebene wurde als Startpunkt die Klassifizierung von (Pinkal 1981)

benutzt. Für die Laufzeit des Projekts haben wir aus dem o.g. Schema wegen besonderer Angemessenheit für das zu analysierende Korpus folgende mögliche Vagheitsindikatoren ausgewählt:

Auf lexikalischer Ebene: Non-Intersectives, Adjektive, Hecken, inexakte Maße, Modalverben (Attitudes), Komplexe Quantoren, Zitiereinleitungen, zeitliche Ausdrücke.

Auf syntaktische Ebene: Subjunktiv-Konstruktionen

Zusätzlich werden Named Entities untersucht: Personen, Zeitangaben, Orte etc. und mit einem entsprechenden Vagheitsgrad versehen ("Konstantinopel" ist nur wahr zwischen 337 und 1930).

Als Vorbereitung wird das Korpus zuerst einer linguistischen Analyse unterzogen, um Lemmas und Wortarten, sowie die Textstruktur (Sätze, Paragraphen) zu markieren. Diese wird dann die Basis für die semi-automatische Annotation von Vagheitsausdrücken (Vertan et al 2017).

Die Annotation von vager Information wird dann in einem ersten Schritt manuell von Fachwissenschaftlern in einem Korpus-Ausschnitt vorgenommen. In einem zweiten Schritt wird versucht diese Annotation automatisch im Korpus zu propagieren. Ein dritter Schritt soll die Ergebnisse von Inferenzen zwischen vagen Ausdrücken erzeugen, um sich nicht dem Vorwurf auszusetzen man schreibe mit einem spezifischen Erkenntnis leitendem Interesse zunächst Annotationen in den Text um sie dann nur wörtlich wieder auszulesen.

Zusammenfassung

Der Beitrag wird die gesamte Systemarchitektur, sowie die einzelnen Schritte zur Annotation von Linguistischer Vagheit illustrieren.

Um dem Wissenschaftler am Ende eine hermeneutische Interpretation zu erlauben, muss ihm zu jedem annotierten Objekt ein Vagheits-Profil sowie Metadaten über Autoren, Genres und Inferenzergebnisse gezeigt werden können.

Hierzu eine Erweiterung von TEI und ein entsprechendes GUI zu entwickeln, sind wichtige Ziele des Projekts.

Außerdem muss die multilinguale Struktur der Cantemir-Texte mit Zitaten aus dem Griechischen, Lateinische und teilweise dem Türkischen annotiert und weiter erforscht werden.

Das Projekt soll zeigen, dass die Einbeziehung von Vagheit und Unschärfe in die Annotierung, in die Inferenz-Komponente und die hermeneutische Interpretationen

durch den Wissenschaftler, einen erheblichen Gewinn an Funktionen und Glaubwürdigkeit für die DH bringt.

Bibliographie

Cantemir, Dimitrie, (1771) Beschreibung der Moldau, Faksimiledruck der Originalausgabe von 1771, Frankfurt und Leipzig

Cantemir, Dimitrie, (1745) Geschichte des osmanischen Reichs nach seinem Anwachsen und Abnehmen, 1745, Herold, Hamburg

Pinkal, Manfred, (1981) Semantische Vagheit: Phänomene und Theorien, Teil I/II. In: Linguistische Berichte Nr. 7/72, Wiesbaden 1980/1981.

Vertan, Cristina / von Hahn, Walther / Dinu, Anca (2017) On the annotation of vague expressions: a case study on Romanian historical texts, Proceedings of the first Workshop on Language Technology for Digital Humanities in Central and (South-) Eastern Europe, in association with RANL 2017, Varna