Linked Open Tafsir
Rekonstruktion der Entstehungsdynamik(en) des Korans mithilfe der Netzwerkmodellierung früher islamischer Überlieferungen

Ahmed, Sajawel
sahmed@em.uni-frankfurt.de
Goethe-Universität Frankfurt am Main, Deutschland

Rehman, Misbahur
rehman@em.uni-frankfurt.de
Goethe-Universität Frankfurt am Main, Deutschland

Tischlik, Joshua
tischlik@em.uni-frankfurt.de
Goethe-Universität Frankfurt am Main, Deutschland

Kruse, Carl
ca.kruse@em.uni-frankfurt.de
Goethe-Universität Frankfurt am Main, Deutschland

Mahmutovic, Edin
mahmutovic@em.uni-frankfurt.de
Goethe-Universität Frankfurt am Main, Deutschland

Özsoy, Ömer
oezsoy@em.uni-frankfurt.de
Goethe-Universität Frankfurt am Main, Deutschland

Das Projekt *Linked Open Tafsir*[1] hat die Erstellung einer online abrufbaren Datenbank früher exegetischer Überlieferungsmaterialien auf Basis des Kommentarwerks von At-Tabari (gest. 310 n. H. / 923 n. Chr.) zum Ziel. Sein Werk *Jami' al-bayan 'an ta'wil ay al-Qur'an* (kurz: Tafsir At-Tabari) kann nach heutigem Kenntnisstand als Sammlung des Großteils aller zu Anfang des 4./10. Jahrhunderts vorliegenden exegetisch relevanten Überlieferungen gelten. In der Datenbank werden die in diesen Überlieferungen enthaltenen Informationen zu historischen Begebenheiten zur Offenbarungszeit sowie den kulturellen, religiösen, sozialen und sprachlichen Rahmenbedingungen der Koranentstehung erfasst. Für die Erfassung der Überlieferungen bzw. Informationen in den Überlieferungen werden mit Hilfe von Künstlicher Intelligenz Daten und Programme insbesondere für das *Named Entity Recognition* entwickelt, die den Erfassungsprozess signifikant beschleunigen. Das Projekt soll eine solide Forschungsgrundlage für Zugänge zur Reflexion der Offenbarungsdynamik(en) des Korans in der frühen Exegese schaffen. Die Datenbank wird insofern "offen" sein, als dass zu einem späteren Zeitpunkt weitere exegetische Kompilationen sowie frühere Überlieferungswerke hinzugefügt werden können. Weiterhin werden die beteiligten WissenschaftlerInnen diese digitalen Zugänge in ihren Forschungsprojekten im Hinblick auf das Islamische Recht, die Systematische Theologie, die Hadithwissenschaft, die Tafsirgeschichte, die Religionspädagogik und dem Maschinellen Lernen für klassische arabische Literatur reflektieren.

Datenbank: Die Erstellung der Datenbank erfolgt in mehreren Schritten: Zunächst sammeln die Projektbeteiligten bereits digital verfügbare Überlieferungen des Kommentarwerks von At-Tabari, um daraus ein erstes digitales Textkorpus zu erstellen. Im Anschluss wird eine Datenbank aufgebaut, die alle in den exegetischen Überlieferungen erhaltenen Informationen über das Offenbarungsumfeld (Mikro-, Makro- und Sprachumfeld), Lesarten (Qira'at), intratextuelle Zusammenhänge (Wiederholung, Querreferenz, Abrogation, Spezifikation etc.) und insbesondere *Named Entities* (NEs: Ort, Zeit, Person etc.) als solche erfasst, markiert, vernetzt und auffindbar macht. Ein weiterer Arbeitsschwerpunkt liegt in der Erschließung der Überlieferungsketten (Isnade) und einzelner TradentInnen. Die Datenbank soll entsprechend verschiedener Forschungsinteressen genutzt werden können: So werden sich beispielsweise alle exegetischen Überlieferungen aus dem Werk Tafsir At-Tabari, welche verschiedene Koranverse mit einer bestimmten Person in Verbindung bringen, in einem einzigen Suchvorgang auffinden lassen. Ebenso werden sich durch die Datenbank unmittelbar Teilkorpora anzeigen lassen, die über dieselben TradentInnen überliefert worden sind. Es ist geplant, die Funktionalität der Datenbank sowie exemplarische Suchmöglichkeiten in Form von YouTube-Tutorials [2] vorzustellen.

Annotation von Named Entities: Unter Named Entity Recognition (NER) versteht man die computerlinguistische Aufgabe, Eigennamen (NE) in Texten zu erkennen (z.B. Mekka, Asien, Tabari, Shia). Solche Eigennamen stehen im Kontrast zu Gattungsnamen, welche eine Klasse von Eigennamen umfassen (z.B. Stadt, Kontinent, Person, Organisation). Technisch gesehen sind für NER zwei Schritte notwendig: Zuerst müssen in einem laufenden Text die Inhaltselemente gefunden werden, die zu einem Eigennamen gehören, danach können diese Eigennamen semantischen Kategorien zugeordnet werden. In unserer aktuellen Annotationsarbeit haben wir aufbauend auf *Guidelines für die Named Entity Recognition* (Benikova 2014; Ahmed 2019) fünf semantische Hauptklassen für klassische arabische Texte unterschieden (Personen, Organisationen, Orte, Zeiten und Andere).

Zurzeit annotiert unser Team aus Arabisten den arabischsprachigen Text, welcher durch einen OCR-Prozess auf historische Manuskripte der Koranexegesis von At-Tabari generiert wurde und uns im XML-basierten TEI-Format vorliegt. Auf Basis dieser digitalen Textsammlung werden von unseren Annotatoren die einzelnen Eigennamen mithilfe des Tools *Oxygen XML Editor* [3] identifiziert, markiert und als NE im TEI-Format abgespeichert. In der *Fig. 1* bekommen wir einen Einblick in der Annotationsumgebung. Wir sehen, dass der Oxygen XML Editor in der Lage ist, den arabischen *Right-to-Left*-Text korrekt darzustellen, die markierten NEs farbig hervorzuheben und insgesamt den Annotatoren einen einfacheren Zugang zum Quelltext zu gewähren.

Abb. 1: Ausschnitt aus dem *Oxygen XML Editor* für die Annotation von *Named Entities* im klassischen arabischen Quelltext *Tafsir At-Tabari*.

Aktueller Stand der Annotationsarbeit und Ausblick: Aktuell haben wir bereits über 30.000 Sätze mit solchen NEs annotieren können, hinzu wird eine weitere solche Menge in der zweiten Phase der Annotation kommen. Zum Abschluss dieser Arbeit werden wir der DH-Fachcommunity den ersten Datensatz überhaupt für die klassische arabische Sprache von solcher Art und Dimension als Open-Source Ressource (Lizenz: CC-BY-4.0) auf *GitHub* bereitstellen, welcher ein ideales Fundament für weiterführende Anwendungen von Sprachmodellen wie *Word2vec* (Mikolov 2013), *LSTM* (Lample 2016; Ahmed 2018), *BERT* (Devlin 2019) sein wird. Begleitend hierzu werden wir eine erste Datenanalyse mit eben diesen Sprachmodellen durchführen und über die Ergebnisse (Precision, Recall und F1-Score) berichten. Unsere innovative, interdisziplinäre Annotationsarbeit legt damit die ersten Bausteine für die Analyse klassischer arabischer Texte mit modernen Verfahren des Maschinellen Lernens, so dass auch der Bereich der Islamwissenschaft und historischen Theologie von der Digitalisierungswelle profitieren kann.

Fußnoten

1. www.linkedopentafsir.de/
2. www.youtube.com/watch?v=LJODcc_Gz50
3. www.oxygenxml.com/

Bibliographie

Ahmed, S. and Mehler, A. (2018): “Resource-size matters: Improving neural named entity recognition with optimized large corpora”, in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 919-924).

Ahmed, S., Stoeckel, M., Driller, C., Pachzelt, A. and Mehler, A. (2019): “BIOfid
Dataset: Publishing a German gold standard for named entity recognition in historical biodiversity literature”, in: *Proceedings of the 23rd Conference on Computational Natural Language Learning* (CoNLL) (pp. 871-880).

Benikova, D., Biemann, C. and Reznicek, M. (2014): “NoSta-D Named Entity Annotation for German: Guidelines and Dataset”, in: *Proceedings of the 9th International Conference on Language Resources and Evaluation* (LREC) (pp. 2524-2531).

Devlin, J., Chang, M.W., Lee, K. and Toutanova, K. (2018): “Bert: Pre-training of deep bidirectional transformers for language understanding”, in: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (NAACL-HTL) (pp. 4171-4186).

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K. and Dyer, C. (2016): ”Neural architectures for named entity recognition”, in: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (NAACL-HTL) (pp. 260-270).

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. and Dean, J. (2013): “Distributed representations of words and phrases and their compositionality”, in: *Advances in neural information processing systems* (NIPS) (pp. 3111-3119).