

Julian Jarosch

DIGITALE LEXIKOGRAFIE MIT HAUSMITTELN

Ein Fallbeispiel zum digitalen Arbeiten ohne Sachmittel

Keywords Tuareg; Citizen Science; Retrodigitalisierung; Digitale Geisteswissenschaften; Onomastik; Linked Open Data

Die lexikografischen Arbeiten des französischen Forschungsreisenden und Mönchs Charles de Foucauld (1858–1916) sind bis heute unübertroffen in der Forschungsgeschichte der Berberlinguistik (Ritter 2009, S. XVI) zur Varietät des Tamahaq (exonymisch: Ahaggar-Tuareg; afroasiatische Familie; gesprochen in der zentralen Sahara, heute im Süden von Algerien), da sie als Produkt von über zehnjähriger immersiver Feldforschung (ca. 1904–1916) entstanden sind. Im Kontrast zu ihrer wissenschaftlichen Bedeutung ist ihre Zugänglichkeit begrenzt: Sie sind überwiegend seit Jahrzehnten nicht mehr im Druck, und in einem Fall bisher nur als Faksimile der handschriftlichen druckfertigen Endfassung des Autors erschienen.

Die zentralen lexikografischen Werke sind ein Eigennamenwörterbuch (Foucauld 1940) und ein umfangreiches allgemeinsprachliches Wörterbuch (Foucauld 1951–1952). Das Eigennamenwörterbuch enthält u. a. Toponyme, Personennamen und Tiernamen. Das allgemeinsprachliche Wörterbuch ist eine detaillierte lexikografische Ressource mit tiefgehenden semantischen Informationen, die in Ansätzen sogar ein kulturelles Lexikon bilden, das eine fast noch vorkoloniale Epoche widerspiegelt. Dieses 2000-seitige zentrale Hauptwerk ist nur als Handschrift-Faksimile publiziert.

Seit 2009 verfolge ich die Erschließung der lexikografischen Werke Foucaulds. Es handelt sich nicht um ein institutionalisiertes oder gefördertes Projekt – der Einblick in die laufenden und abgeschlossenen Arbeiten fungiert deshalb als Beispiel für Forschung mit öffentlich und kostenfrei verfügbaren Werkzeugen und Infrastrukturen. Die Auswahl der Teilbereiche deckt dabei einschlägige Eckpunkte des digitalen Forschungsprozesses ab: Text-Retrodigitalisierung, digitale Editorik, semantische Anreicherung und wissenschaftliche Auswertung.

Für die Digitalisierung des handschriftlichen Textes verwende ich automatische Handschrifterkennung in der Software Transkribus (<https://readcoop.eu/transkribus/>). Dieses Werkzeug erlaubt es, neuronale Netze auf die Erkennung beliebiger Handschriften zu trainieren. Voraussetzung ist, dass transkribierte und korrigierte Textseiten als Trainingsmaterial vorliegen, was hier durch die manuelle Transkription von mehreren hundert Seiten umgesetzt wurde. Diese Methode kann im vorliegenden Anwendungsfall auch die Mischung von Sprachen, Schriftsystemen und Schreibrichtungen adäquat berücksichtigen.

Die Edition und langfristige Publikation der Wörterbuchtexte erfolgt in Wikisource, einem Schwesterprojekt der Wikipedia. Wikisource steht als kostenfreie öffentliche Infrastruktur zur Textdigitalisierung zur Verfügung und bietet als Arbeitsplattform wesentliche und anpassbare Werkzeuge wie Sonderzeichen-Tastaturen und Templates für wiederkehrende Text- oder Layoutbausteine. Darüber hinaus handelt es sich um eine offene, Crowdsourcing-basierte Plattform. Dies bedeutet, dass die digitalisierten Texte nicht in einer End-

fassung fossilisieren, sondern dass zu jedem Zeitpunkt die Genauigkeit der Transkription gegen die Originale geprüft und, wenn nötig, weiter verbessert werden kann – ohne Zugangsschwellen.

Ein Forschungszweig im hier vorgestellten Projektkontext ist die semantische Anreicherung des Toponyme-Gazetteers im Eigennamenwörterbuch durch Georeferenzierung – also die Zuordnung der im Namenbuch (mit groben geografischen Informationen) genannten Orte zu präzisen geoinformatischen Objekten. Zur eindeutigen und dauerhaften Identifikation der Orte wird die offene geografische Normdatenbank GeoNames (<http://www.geonames.org/>) verwendet, die Geografika unveränderliche Identifikatoren zuweist und langfristig verfügbar vorhält. Die punktförmig groben Verortungen von GeoNames ermöglichen im nächsten Schritt die Zuordnung der Wörterbucheinträge zu geografischen Objekten in der Geo-Datenbank und Online-Karte OpenStreetMap (<https://www.openstreetmap.org/>), was eine deutlich genauere und adäquatere Repräsentation von linienförmigen und flächigen Orten ermöglicht (beispielsweise Trockenflüsse und Gebirge). OpenStreetMap ist ebenfalls ein durch Crowdsourcing und Nutzerbeiträge getragenes Projekt und kann deshalb in der Untersuchungsregion um die benötigten Daten ergänzt werden. Die Zusammenführung der drei genannten Datensätze – der Gazetteer von Foucauld, die eindeutigen Identifizierungen in GeoNames, und die geografischen Daten in OpenStreetMap – können in eine kartografische Darstellung der Ortsnamenbuch-Informationen resultieren.

Auf der linguistischen Ebene können die Toponyme auch als beispielhafte Auswertung der retrodigitalisierten Informationen zur Beantwortung einer wissenschaftlichen Fragestellung dienen. Während für Eigennamen eine referentielle Einzigkeit als gesicherte wissenschaftliche Erkenntnis etabliert ist (Nübling/Fahlbusch/Heuser 2012, S. 17–20), scheint das Ortsnamenbuch von Foucauld diesem Monoreferenzprinzip zu widersprechen, indem es knapp über ein Viertel der Ortsnamen mehr als einen Referenten zuordnet. Eine Frequenz- und areale Verteilungsanalyse der Toponyme, ermöglicht durch die Digitalisierung der Daten, erlaubt die Interpretation des Phänomens als Funktion der nomadischen Lebensweise in der ariden Landschaft: Ausgeprägte Ortskenntnisse sind dort überlebensnotwendig, aber ein unbegrenzt umfangreiches Onomastikon ist mit der menschlichen Gedächtniskapazität unvereinbar; die Ortsnamen für mehrere geografische Objekte können also als mnemonisches Mittel zur Gedächtnisökonomie eingeschätzt werden (Jarosch/Späth 2021).

Das Projekt berührt also an mehreren Punkten gesellschaftliche Aspekte: Als überwiegend crowdsourcing-offenes Projekt ist es partizipatorisch angelegt für citizen scientists; die Wörterbücher enthalten eine Vielfalt an wissenschaftlich relevanten, historisch-kulturellen Einblicken in eine forschungsperspektivisch marginalisierte Sprechergemeinschaft; und als erste maßgebliche digitale lexikografische Ressource zum Tamahaq hat es das Potenzial, sprachpolitisch rezipiert zu werden und eine Rolle im Sprach- und Kulturerhalt zu spielen.

Literatur

Foucauld, C. de (1940): Dictionnaire abrégé touareg-français de noms propres. Dialecte de l'Ahaggar. Paris: Larose. https://fr.wikisource.org/wiki/Dictionnaire_abrégé_touareg-français_de_noms_propres (Stand: 04-04-2022).

Foucauld, C. de (1951–1952): Dictionnaire touareg – français. Dialecte de l'Ahaggar. 4 Bände. Paris: Imprimerie Nationale de France. https://fr.wikisource.org/wiki/Dictionnaire_touareg_-_français (Stand: 04-04-2022).

Jarosch, J./Späth, L. (2021): Toponyme einer nomadischen Gesellschaft – Orientierung in einer ariden Landschaft. In: Dräger, K./Heuser, R./Prinz, M. (Hg.): Toponyme. Berlin/New York, S. 87–108. <https://doi.org/10.1515/9783110721140-005>.

Nübling, D./Fahlbusch, F./Heuser, R. (2012): Namen. Eine Einführung in die Onomastik. Tübingen.

Ritter, H. (2009): Wörterbuch zur Sprache und Kultur der Twareg. Band I. Wiesbaden.

Kontaktinformationen

Julian Jarosch

Akademie der Wissenschaften und der Literatur Mainz/Johannes Gutenberg-Universität Mainz
julian.jarosch@adwmainz.de