# Vom Digitalisat zur Ressource: Der Workflow zum Deutschen Wortatlas

**Kunzmann, Markus**
markus.kunzmann@oeaw.ac.at
Österreichische Akademie der Wissenschaften (ÖAW), Österreich
ORCID: 0000-0003-3387-1542

In der Sprachwissenschaft bilden Sprachkorpora eine effiziente Möglichkeit, Fragen zu unterschiedlichen linguistischen Strukturebenen auf empirischer Grundlage zu beantworten. Große Korpora wie das Deutsche Referenzkorpus (DeReKo)[1] (Kupietz et al. 2023) oder das Austrian Media Corpus (AMC)[2] (Ransmayer & Pirker 2023) repräsentieren dabei meist einen gegenwärtigen oder zumindest jüngeren Sprachstand der Deutschen Sprache. Korpora, die einen älteren Sprachstand widerspiegeln, z.B. das Referenzkorpus Frühneuhochdeutsch (ReF)[3] (Herbers et al. 2021), beziehen große Teile ihrer Texte ebenfalls aus gedruckten Werken oder Handschriften, die bereits ediert wurden. Texte, die bislang nur in handschriftlicher Form vorliegen und die noch nicht editorisch verarbeitet wurden, sind aufgrund der zeitintensiven Erschließung oft nicht Teil dieser Korpora. Dies betrifft auch den Großteil der indirekten Erhebungen zu Dialekten, die vor allem zu Beginn des 20. Jahrhunderts durchgeführt und meist handschriftlich von den Gewährspersonen ausgefüllt wurden. Beispiele hierfür wären die Erhebungen zum Deutschen Sprachatlas (Wenker et al. 1927), zum Deutschen Wortatlas (Mitzka et al. 1980) oder die sogenannten Maurer-Fragebögen[4].

Erst mit der Entwicklung KI-gestützter Verfahren zur Handschrifterkennung (Graves & Schmidhuber 2007) ist es auch möglich, größere, bislang hauptsächlich handschriftlich vorliegende Textbestände, im Volltext zu digitalisieren. Mit Softwarelösungen wie Transkribus[5] oder eScriptorium[6] wurde diese Möglichkeit aufgrund deren vergleichsweise niederschwelligen Anwendung noch zusätzlich beschleunigt. Ein Beispiel für hauptsächlich handschriftlich vorliegende Bestände, die ohne *Handwritten Text Recognition* (HTR) in diesem Ausmaß nicht erschließbar gewesen wären, sind die zum bereits erwähnten Deutschen Wortatlas (DWA).

Der DWA ist ein Sprachatlas zur deutschsprachigen Lexik des deutschen Sprachraums mit Ausnahme der Schweiz, der in den Jahren zwischen 1951 und 1980 in 22 Bänden herausgegeben wurde. Grundlage dafür waren indirekte Erhebungen aus den Jahren 1939 bis 1942, die an ca. 50.000 Schulorte verschickt wurden. Dabei sollten 188 Einzelwörter und zwölf Sätze durch Lehrerinnen bzw. Schülerinnen und Schüler in den eigenen Dialekt übersetzt werden (Mitzka 1938). Da den Probandinnen und Probanden i.d.R. nur ihr schulisch erworbenes Alphabet ohne die Möglichkeit, lautliche Eigenheiten zu annotieren zur Verfügung stand, eignen sich die Daten des DWA in erster Linie als Quelle für den Wortschatz und weniger für phonetisch-phonologische Fragestellungen. Bis heute können die Erhebungen zum DWA als die größten lexikalisch orientierten Erhebungen gelten, die zum Deutschen durchgeführt wurden. Die Karten des DWA präsentieren sich als sogenannte Punktsymbolkarten, die nach einem meist interessenspezifischen Prinzip die einzelnen Belege einigen übergeordneten Worttypen zuordnen. Dieses in der Sprachkartographie lange verankerte Prinzip hat zum Vorteil, dass es einen raschen Überblick zu arealbildenden Merkmalen des Sprachmaterials bietet. Allerdings ist dies meist auf den Aspekt beschränkt, unter dem die Karte erstellt worden ist. Es fehlen somit die Sprachdaten selbst, so wie sie in den originalen Erhebungsbögen zu finden sind.

Im Projekt „DWA Österreich Pilotstudie“[7] (Kunzmann 2023) wurde eine Volltexterschließung hinsichtlich ihrer Machbarkeit untersucht. Das Pilotprojekt wurde 2022/2023 am Austrian Centre for Digital Humanities and Cultural Heritage (ACDH-CH) der Österreichischen Akademie der Wissenschaften (ÖAW) in Zusammenarbeit mit der Universität Wien und dem Forschungszentrum Deutscher Sprachatlas (DSA) der Universität Marburg durchgeführt. Dabei wurde ein Digitalisierungs-Workflow erarbeitet und auf der Grundlage von 200 Bögen aus dem Gebiet des heutigen Österreich getestet. Der Workflow umfasst in erster Linie die Schritte der Vorbereitung der Digitalisate[8] für die automatische Transkription in Transkribus, die Korrektur der Ergebnisse zur Schaffung eines Ground-Truth-Datenbestandes sowie den automatischen Export in XML/TEI zur weiteren Nutzung der Ergebnisse. Hierfür bietet die Plattform auch die Möglichkeit, Dokumente sowohl mit Struktur- als auch Text-Tags zu versehen Anwendungsseitig wird die Character Error Rate (CER) bzw. die Word Error Rate (WER) jeweils auf das Trainingsset selbst oder ein eigenes Validierungsset angeboten.

Das Poster soll die einzelnen Bearbeitungsschritte des Digitalisierungsworkflows sowie die weitere Verarbeitung und Nutzung der Daten umfassen. Ein wesentlicher Bestandteil soll auch die Gegenüberstellung des zeitlichen Aufwands von manueller Transkription, automatischer Transkription auf der Grundlage bereits vorhandener HTR-Modelle sowie die Anwendung quellenspezifischer HTR-Modelle bilden. Dabei soll insbesondere der Ressourcenaufwand zum Aufbau solcher Modelle kritisch beleuchtet werden. Die Verarbeitung profitiert einerseits von der Einheitlichkeit des Aufbaus der Quelldokumente, andererseits sind sie im Hinblick auf das Schriftbild und den Inhalt äußerst heterogen, was den praktikablen Einsatz bereits vorhandener größerer Texterkennungsmodelle ausschließt. Diesbezüglich soll auch diskutiert werden, welche Einschränkungen die CER als Kennzahl für die Güte eines HTR-Modells mit sich bringt.

# Fußnoten

1. Deutsche Referenzkorpus (DeReKo) auf den Seiten des Leibniz-Institut für Deutsche Sprache (IDS): https://www.ids-mannheim.de/digspra/kl/projekte/korpora/
2. Austrian Media Corpus (AMC) auf den Seiten des Austrian Centre for Digital Humanities and Cultural Heritage (ACDH-CH): https://amc.acdh.oeaw.ac.at/
3. Referenzkorpus Frühneuhochdeutsch (ReF) auf den Seiten der Ruhr-Universität Bochum: https://www.linguistics.rub.de/ref/
4. Informationen zu den sog. Maurer-Fragebögen auf den Seiten des Bayerischen Wörterbuchs (BWB) der Bayerischen Akademie der Wissenschaften (BAdW): https://bwb.badw.de/materialsammlung/die-erhebungen.html#c4596
5. Transkribus-Plattform auf den Seiten der READ COOP SCE: https://www.transkribus.org/de
6. eScriptorium: https://escriptorium.inria.fr/
7. Infoseite zum Projekt „DWA Österreich Pilotstudie“ auf den Seiten des Austrian Centre for Digital Humanities and Cultural Heritage (ACDH-CH): https://www.oeaw.ac.at/de/acdh/projects/dwa-oesterreich
8. Die Digitalisate der Erhebungsbögen zum Deutschen Wortatlas werden auf dem Repositorium des Forschungszentrums Deutscher Sprachatlas LinguRep ( https://hdl.handle.net/20.500.14450/429) der Philipps-Universität Marburg zur Verfügung gestellt.

# Bibliographie

**Graves, Alex & Jürgen Schmidhuber**. 2009. Offline Handwriting Recognition with Multidimensional Recurrent Neural Networks. *Advances in Neural Information Processing Systems* 22, S. 545–552. URL: https://people.idsia.ch/~juergen/nips2009.pdf (Stand: 24.07.2024)

**Herbers, Birgit, Sylwia Kösser, Ilka Lemke, Ulrich Wenner, Juliane Berger, Sarah Kwekkeboom und Frauke Thielert**. 2021. Dokumentation zum Referenzkorpus Frühneuhochdeutsch und Referenzkorpus Deutsche Inschriften. *Bochumer Linguistische Arbeitsberichte* 24. URL: https://linguistics.rub.de/forschung/arbeitsberichte/24.pdf (Stand: 24.07.2024)

**Kunzmann, Markus**. 2023. AI-supported indexing of handwritten dialect lexis: The pilot study "DWA Austria" as a case study. In: Baillot, Anne, Toma Tasovac, Walter Scholger und Georg Vogeler (Hrsg.). *Annual International Conference of the Alliance of Digital Humanities Organizations Graz, Austria, July 10-14, 2023, Book of Abstracts*, S. 80–81. https://zenodo.org/doi/10.5281/zenodo.7961821

**Kupietz, Marc, Harald Lüngen und Nils Diewald**. 2023. Das Gesamtkonzept des Deutschen Referenzkorpus DeReKo. In: Deppermann, Arnulf, Christian Fandrych, Marc Kupietz und Thomas Schmidt (Hrsg.). *Korpora in der germanistischen Sprachwissenschaft: Mündlich, schriftlich, multimedial*. Berlin, Boston: De Gruyter, S. 1–28. https://doi.org/10.1515/9783111085708-002

**Mitzka, Walther** .1938. Der Deutsche Wortatlas, *Zeitschrift für Mundartforschung*, 14/1, S. 40–55.