# Automatisierte Extraktion und Klassifikation von Variantenschreibungen historischer Berufsbezeichnungen in seriellen Quellen des 16. bis 20. Jahrhunderts

**Moeller, Katrin**
katrin.moeller@geschichte.uni-halle.de
Martin-Luther-Universität Halle-Wittenberg, Germany

**Goldberg, Jan Michael**
jan.goldberg@wiwi.uni-halle.de
Martin-Luther-Universität Halle-Wittenberg, Germany

## Einleitung

Berufsangaben kommen in sehr vielen historischen Quellen vor, besonders in seriellen Quellen wie Kirchenbüchern, Adressbüchern, Mieter-, Einwohner- und Bürgerlisten etc. Für eine Vielzahl von Forschungsgebieten bildet nicht nur die Standardisierung, sondern vor allem eine ordnende Klassifikation dieser Nennungen eine zentrale Voraussetzung zur tatsächlichen Analyse von Berufen nach verschiedenen möglichen Ordnungsprinzipien. Einen Ansatz bildet die „Ontologie der historischen Amts- und Berufsbezeichnungen (OhdAB)“ (Moeller 2019, Moeller 2021), die Amts- und Berufstätigkeit des 16. bis 20. Jahrhunderts klassifiziert. Der Vortag beschäftigt sich mit der automatisierten Zuordnung von Variantenschreibungen von Amts- und Berufsbezeichnungen zu den Berufsgattungsnamen dieses Klassifikationsansatzes und damit zu einer verknüpften Klassifikation nach Tätigkeitskonzepten und Anforderungsniveaus, die auf der Methodik der Klassifikation der Berufe 2010/2020 aufbaut (Bundesagentur für Arbeit 2021). Dabei bleibt es unberücksichtigt, ob diese Varianten aus fehlerhaften Lese- und Schreibprozessen durch Mensch oder Maschine bzw. aus Variantenschreibungen der Quellen resultieren. Im Mittelpunkt des Beitrags steht der Algorithmus zur Identifizierung von Berufsbezeichnungen in strukturierten Quellen. Entwickelt wurde eine Vorgehensweise des Machine Learnings zur Erkennung von Variantenschreibungen, die im Vortrag vorgestellt wird. Diese besteht aus einem komplexen Workflow eines automatisierten Preprocessings zur Identifizierung bzw. Separierung der eigentlichen Berufsangaben und einer auf einem Algorithmus beruhenden Zuordnung unbekannter Varianten zur Klassifikation. Dieser Algorithmus wurde auf der Basis bereits zur Klassifikation (OhdAB) zugeordneter Varianten entwickelt und trainiert. Am Beispiel eines unbereinigten und stark heterogenen Datensatzes des Vereins für Computergenealogie (Verein für Computergenealogie 2021) wurde eine Erkennungsrate von 75 Prozent der Berufsangaben ermittelt, wobei nur fünf Prozent fehlerhaften Zuordnungen zu identifizieren sind.

## Berufsangaben in genealogischen Quellen, Preprocessing-Workflow

Amts- und Berufsangaben kommen in halbstrukturierter Form in zahlreichen historischen und modernen Quellen vor. Das Auslesen dieser Information aus einer begrenzten und bereits vorstrukturierten Textmenge bildet besondere Anforderungen an Verfahren des Natural Text Processings ab, da hier andere sprachanalytische Analyseformen zur Anwendung kommen (können) als in Volltexten. Demgegenüber ist die Zahl der Varianten besonders bei dieser Textsorte nicht nur durch Schreibvarianten der Quellen, sondern durch zahlreiche maschinelle Erhebungsverfahren zunehmend auch von OCR- oder HTR-Erkennungsfehlern und vor allem durch Abkürzungen geprägt. Zudem stammen viele maschinenlesbare Massendaten aus der Community der Citizien Sciences, die durch zusätzliche Aufnahme- und Eintragungsbesonderheiten gekennzeichnet sind. Der im Beispiel verwendete Testdatensatz repräsentiert einen besonders stark verunreinigten Datensatz des Vereins für Computergenealogie.

Der wesentliche Vorteil dieser Quellen speist sich jedoch aus der bereits entitätsspezifisch strukturierten Datenmenge, da die aufzeichnenden Personen der Vergangenheit spezifische Vorstellungen von Amt, Beruf bzw. „Berufsstand“ wiedergaben. Dennoch gibt es auch in diesen Quellen eine begrenzte Menge weiterer Informationen, die zum Teil historische Auffassungen vom Berufsstand (z. B. Verwandtschaftsbeziehungen von Frauen und Kindern, Familienstand, Renten- und Altenteilbezüge, Kirchen-, Amts- und Ehrenvorstände etc.) zum Teil aber auch weniger reflektierte Aufnahmepraktiken oder fehlerhafte Einträge wiedergeben.

Keineswegs können solche Angaben in historischen Quellen jedoch wie in der Problemklassifikation nach Rahm und Do (Rahm / Do 2000: 3f.). lediglich als Einquellenprobleme auf einem Level einzelner Instanzen (Berufsangabe) gekennzeichnet und aus der Analyse ausgeschlossen werden. Wie gezeigt, ist für historische Daten dagegen ein kontextualisierender Begriff des Berufsstandes wichtig, der fehlerhafte Einträge erst nach der Zuordnung und Klassifikation bereinigt. Die Angabe des Rechtsstatus oder Familienstandes kann eine Person in ihrem Stand ebenso adäquat beschreiben, während eine Ortsangabe nur eine in das falsche Datenfeld eingetragene Information repräsentieren kann. Dies berücksichtigt die historische Klassifikation OhdAB, indem sie heutige Berufsgegenstände von anderen Entitäten separiert, beide Formen jedoch ordnet und analysiert. Zur Lösung dieser qualitativen Probleme schlagen Müller und Freytag (Müller / Freytag 2003: S. 10-13) einen vierstufigen Prozess der Datenbereinigung vor. An dessen Beginn steht ein Datenaudit ( *data auditing* ), in welchem die Daten geparst und analysiert werden. Dadurch werden syntaktische Anomalien erkannt, die es anschließend zu bearbeiten gilt. Dazu wird in einem zweiten Schritt der Ablauf der Datenbereinigung spezifiziert ( *workflow spezification* ). Dabei kann die Behebung syntaktischer Fehler im Nachhinein wiederum andere Anomalien sichtbar machen. Die nachfolgende Durchführung der Datenbereinigung ( *workflow execution* ) steht im Konflikt zwischen einer möglichst passenden Korrektur und einer akzeptablen Laufzeit. Manuelle Nacharbeit ist zu vermeiden, da diese Ressourcen binden. Eine nicht-automatisierte Kontrolle findet allerdings in einem vierten Schritt statt ( *post-processing and controlling* ). Hierfür wird mit dem Beitrag ein konkreter Work-

flow zur Extraktion von Berufen und Zuweisung der Berufsklassifikation vorgestellt.

Diese Datenbereinigung und das Preprocessing bleibt selbst bei den strukturierten Angaben komplex, zeigt aber durchaus Verbesserungspotential beim Datenmatching. Spezifische Problemlagen der Berufsbezeichnungen boten neben den mehr oder weniger spezifischen Abkürzungen vor allem die Angaben von mehreren Amts- und Berufsbezeichnungen in verschiedenen Sinnkonstruktionen. So können Berufsangaben immer wieder „paarig“ genannt werden, wie der Beruf des Gold- und Silberschmieds und damit einen gemeinsamen Berufsgattungsnamen repräsentieren oder eben auch Reihungen von verschiedenen Berufsnamen enthalten, die nacheinander klassifiziert werden müssen. Gleichzeitig wurden temporale Hinzufügungen, Präzisierungen zum konkreten Berufs- oder Arbeitsort, Firmen- oder Einheitsangaben, Besitzinformationen etc. oder fremdsprachliche Angaben identifiziert. Durch das Konzept des Berufsstandes spielen zudem Angaben zum Familienstand, zur Rolle innerhalb der Familie (arbeitende Ehefrau, Witwe oder Kinder), Rechtsinformationen sowie Standestitel eine wichtige Rolle. Daneben gibt es über die zahlreichen ergänzenden Informationen hinaus immer wieder auch falsch angeordnete Entitäten (Namen oder Ortsangaben, weitere Eigennamen oder auch Quellenbezeichnungen).

Aufgrund der qualitativen Datenanalyse wurde ein Kanon von Separatoren und Trennzeichen ermittelt, der verschiedene Informationsketten „anzeigt“, markiert und je nach Qualität auszeichnet, separiert oder löscht. So wurden bspw. über lokale Präpositionen Ortsangaben, über temporale Präpositionen zeitliche Angaben zum Beruf separiert und normiert. Andererseits bildete ein Vokabular zu verschiedenen Formen von Verwandtschaftsbezeichnungen oder des Familienstandes die Grundlage zur separierten Definition der Familienrolle, die für die Einordnung der eigentlichen Ausübung von Tätigkeiten zentrale Informationen abbildet. Im Vortrag sollen die entsprechenden Möglichkeiten dazu kurz systematisch skizziert und in ihrer Funktionalität dargestellt werden.

# Algorithmus zur Variantenzuordnung

Da Berufsangaben Strings im Sinne einer semantischen Zeichenkette darstellen, können String-Matching-Algorithmen zur Erkennung einer unscharfen Übereinstimmung auf sie angewendet werden. Die Ähnlichkeit von Strings kann über verschiedene Maße ausgedrückt werden. In der historischen Linguistik stellt die Levenshtein-Distanz eine geeignete Möglichkeit dar, die infrage kommende Beziehung zwischen Wörtern aufzuzeigen. Die Herausforderung, zwei Schreibvarianten desselben Wortes zu erkennen, ist ähnlich gelagert wie die Erkennung einer möglichen linguistischen Verwandtschaft zwischen zwei Wörtern.

Zunächst sollen möglichst viele Berufsangaben den richtigen Entitäten, im Weiteren „Klassen“, zugeordnet werden (True Positiv = TP). Ein Berufsgattungsname stellt dabei eine Klasse dar; die bekannten Schreibweisen (Varianten) wiederum sind die Eigenschaften. Eine Übersicht über die verwendeten Begrifflichkeiten ist, insbesondere für die multiple Verwendung der Klassifizierung / Klassifikation, in Abbildung 1 ersichtlich.

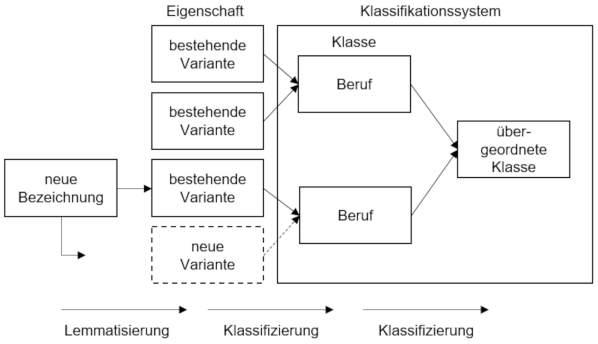


Abb. 1: Begriffe und Zusammenhänge des Algorithmus. [Goldberg / Moeller 2021]

Eine Erhöhung der TP-klassifizierten allein geht jedoch oftmals auch mit der Erhöhung von FP-Klassifizierungen (False Positiv) einher. Aus diesem Grund wird nicht die Anzahl der TP-Klassifizierungen optimiert, sondern das $F_1$-Maß als Ausweis dieser falsch zugeordneten Begriffe. Zudem soll die Klassifizierung automatisch geschehen; eine manuelle Überprüfung des Ergebnisses geschieht nicht. Das ist notwendig, um große Datenbestände in einer überschaubaren Zeit klassifizieren zu können. Da der Algorithmus insbesondere auf große Listen von Berufsangaben Anwendung finden soll, ist dessen Effizienz und somit die Laufzeit zu beachten. Der Algorithmus ist in einem Programmcode (Python basiert) umgesetzt worden, der in weiteren Applikationen eingebunden werden kann.

Nach der Bereinigung sind den Berufsangaben trotzdem noch keine Berufsgattungsnamen der OhdAB-Konkordanz zugeordnet. Die notwendige Zuordnung geschieht auf Basis der Eigenschaften der bestehenden Klassen. Darum findet ein Abgleich mit den vorhandenen Varianten der OhdAB statt. Eine Berufsangabe soll der Klasse zugeordnet werden, deren Zugehörigkeit am wahrscheinlichsten ist. Die Ähnlichkeit einer Berufsangabe zu den Eigenschaften (bestehende Varianten) einer Klasse (Beruf) wird dabei als Indikator für die Wahrscheinlichkeit einer korrekten Zuordnung (Normierung/Lemmatisierung[1] ) genutzt. Diese kann über einen Vergleich der Zeichenketten ermittelt werden. Jedoch muss nicht zwingend eine Lemmatisierung stattfindet: Wenn die Ähnlichkeit zu jeder Klasse so gering ist, dass eine korrekte Zuordnung unwahrscheinlich ist, kann kein Pendant gefunden werden.

Zeichenketten können auf verschiedene Arten verglichen werden. Kirby et al. empfehlen für die weitere Forschung eine Variation von verschiedenen Vergleichsmethoden (Kirby, 2015, S. 58). Dadurch, dass die Variante einer Normschreibweise der Konkordanz zugeordnet ist, ist auch ihre Zuordnung zu einer Berufsgattung der OhdAB eindeutig. Besteht keine Übereinstimmung mit einer Variante, so ist eine teilweise Übereinstimmung zu überprüfen. Daher wurden für die Entwicklung des Algorithmus eine Reihe verschiedener Ansätze ausgetestet, die einerseits eine 1:1 Zuordnung (unter Verzicht von Groß- und Kleinschreibung) sowie verschiedene Variationen der Levenshtein-Distanz und weiterer Iterationsschritte umfassen.

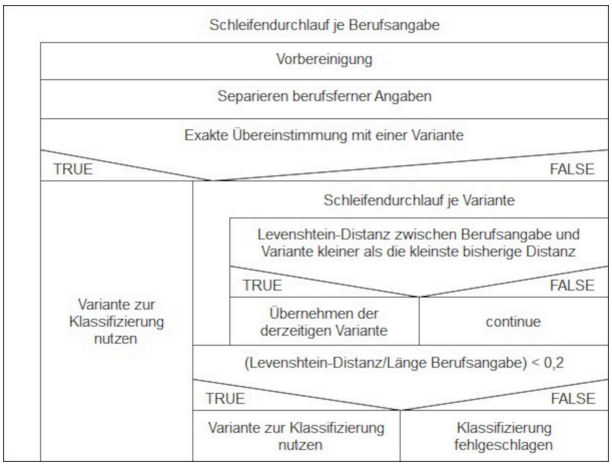


Abb. 2: Algorithmus, dargestellt in einem Nassi-Shneiderman-Diagramm. [Goldberg / Moeller 2021]

Im Vortrag möchten wir unser Vorgehen wie auch die ermittelten Testwerte vorstellen und genau erläutern (was hier aufgrund der verfügbaren Platzmenge nicht stattfinden kann). Zusammengefasst wird das $F_1$-Maß optimiert, wenn eine relative Levenshtein-Distanz gewählt wird, Abkürzungen erweitert werden und erlernte neue Varianten im Anschluss nochmal mit allen Daten verglichen werden, die nicht lemmatisiert werden konnten. Der letztgenannte Aspekt des teil-maschinellen Lernens führt dazu, dass neue Varianten stetig hinzukommen können. Der Algorithmus wird auf die Testdaten (220.000 Berufsangaben) angewendet. 64 Prozent der beruflichen Bezeichnungen können direkt lemmatisiert und einer bestehenden Variante zugeordnet werden. Bei den weiteren wird eine Ähnlichkeitsanalyse durchgeführt, die in vier Prozent der Fälle ein Ergebnis erbringt. Insgesamt hat die Ähnlichkeitsanalyse nur eine vergleichsweise geringe Auswirkung. Relevanter ist der Einsatz der Bereinigung: Wird diese gänzlich ausgelassen, so können nur 58 Prozent der Angaben direkt in den Varianten erkannt werden .

| | Direkt gefunden | Ähnlichkeitsanalyse | Nicht gefunden | Leere Bezeichnungen |
|---|---|---|---|---|
| mit Bereinigung (229.699 Angaben) | | | | |
| Anzahl | 147.781 | 9.674 | 68.955 | 3.259 |
| Anteil | 64,35 % | 4,21 % | 30,02 % | 1,42 % |
| ohne Bereinigung (229.669 Angaben) | | | | |
| Anzahl | 131.064 | 9.160 | 86.344 | 3.101 |
| Anteil | 58,07 % | 3,89 % | 37,59 % | 1,35 % |

Tab. 1: Vergleich des Effektes der Bereinigung auf die Erkennung. [Goldberg / Moeller 2021]

Die durch die Ähnlichkeitsanalyse neu zugeordneten Berufsangaben können in der Variantenliste ergänzt werden. Dieses kann geschehen, indem die neuen Treffer direkt nach Erkennung in die Menge der Varianten eingehen oder aber alle nicht erkannten Bezeichnungen anschließend mit allen Treffern abgeglichen werden. Letzteres ist mit mehreren Iterationen denkbar. Hierbei zeigt sich, dass die nachfolgende, iterative Verarbeitung ein besseres Ergebnis in Bezug auf das $F_1$-Maß ergibt als die kontinuierliche Ergänzung (siehe Tabelle 2). Dabei ist der Lerneffekt größer, je mehr Berufsangaben verarbeitet werden, da die Chance steigt, dass eine ähnliche Bezeichnung auftritt. Bei einem Durchlauf mit jeder zehnten Datei wird noch keine zusätzliche Erkennung erreicht. Allerdings werden auch bei einer Verarbeitung aller Daten nur weitere 0,01 Prozent der Berufsangaben lemmatisiert. Dieses ist darauf zurückzuführen, dass bereits sehr viele Schreibversionen in den zugrundeliegenden Varianten der OhdAB abgedeckt sind (momentan ca. 200.000). Bei einer willkürlichen Halbierung der ursprünglichen Varianten steigt der Anteil der so zusätzlich erkannten Angaben deutlich um rund 9 Prozent (von 3,10 Prozent auf 12,01 Prozent). Werden diese lemmatisierten Varianten in einem zweiten Durchlauf zur Gesamtzahl der Varianten ergänzt, können weitere Berufsbezeichnungen lemmatisiert werden. Die TP-Rate jedoch ist etwas niedriger. Eine hohe FP-Rate in der ersten Ähnlichkeitserkennung führt tendenziell zu einer Fortführung von Fehlern.

| | Direkt gefunden | Ähnlichkeitsanalyse | Nicht gefunden | Leere Bezeichnungen |
|---|---|---|---|---|
| mit Bereinigung (229.699 Angaben) | | | | |
| Anzahl | 147.781 | 9.674 | 68.955 | 3.259 |
| Anteil | 64,35 % | 4,21 % | 30,02 % | 1,42 % |
| ohne Bereinigung (229.669 Angaben) | | | | |
| Anzahl | 131.064 | 9.160 | 86.344 | 3.101 |
| Anteil | 58,07 % | 3,89 % | 37,59 % | 1,35 % |

Tab. 2: Vergleich der Klassifikation unter Halbierung der zugrundeliegenden Berufsvarianten der OhdAB. [Goldberg / Moeller 2021]

Durch den Algorithmus – und dessen programmtechnische Umsetzung – wird in der Folge eine automatisierte Lösung zur Lemmatisierung deutschsprachiger Berufsangaben geboten. Mittels Variation der Ähnlichkeitsanalyse konnte zwar formal kein Optimierungsproblem gelöst werden; es hat sich aber gezeigt welche Faktoren das F 1 -Maß verschlechtern und welche es verbessern. Zudem ist es durch den Algorithmus möglich, berufsferne Angaben von der eigentlichen Bezeichnung des Berufs zu separieren.

# Fußnoten

1. Der Begriff der Lemmatisierung wird hier als Zuweisung einer Normschreibung zu verschiedenen Schreibvarianten des gleichen Berufsgattungsnamens verstanden.

# Bibliographie

**Bundesagentur für Arbeit** (2021): *Klassifikationen der Berufe - Statistik der Bundesagentur für Arbeit*. Nürnberg. [online].

**Bundesagentur für Arbeit** (2011): *Klassifikation der Berufe*. Nürnberg 2010. Bd 1 (2011): *Systematischer und alphabetischer Teil mit Erläuterungen.*

**Church of Jesus Christ of Latter-day Saints** (2019): *The GEDCOM Standard*. Release 5.5.1. 2019.

**Theresa Cosca / Alissa Emmel** (2010): "Revising the Standard Occupational Classification system for 2010". In: *Monthly labor review* 133, 32–41. PDF. [online] 320603628.

**Jyldyz Djumalieva / Antonio Lima / Cath Sleeman** (2018): *Classifying Occupations According to Their Skill Requirements in Job Advertisements.* [online].

**Hyukjun Gweon / Matthias Schonlau / Lars Kaczmirek / Michael Blohm / Stefan Steiner** (2017): "Three Methods for Occupation Coding Based on Statistical Learning". In: *Journal of Official Statistics* 33, H. 1, 101–122. DOI: 10.1515/jos-2017-0006 130422746.

**Jan Michael Goldberg / Katrin Moeller** (erscheint 2022): "Automatisierte Identifikation und Lemmatisierung historischer Berufsbezeichnungen in deutschsprachigen Datenbeständen", in: *Zeitschrift für digitale Geisteswissenschaften* (im Druck).

**J. Tuomas Harviainen / Bo-Christer Björk** (2018): "Genealogy, GEDCOM, and popularity implications". In: *Informaatiotutkimus* 37, H. 3, S. 4–14. Artikel vom 29.10.2018. DOI: 10.23978/inf.76066 366701630

**Graham Kirby / Jamie Carson / Fraser Dunlop / Chris Dibben / Alan Dearle / Lee Williamson / Eilidh Garrett / Alice Reid**: "Automatic Methods for Coding Historical Occupation Descriptions to Standard Classification". In: *Population Reconstruction*, Hg. von Gerrit Bloothooft / Peter Christen / Kees Mandemakers / Marijn Schraagen. Cham / Heidelberg u.a. 2015, S. 43-60.

**Thomas Krause** (2012): *Entwurf und Implementierung einer effizienten Dublettenerkennung für große Adressbestände*. Köln. URN: urn:nbn:de:hbz:832-epub-3667.

**Marco H. D. van Leeuwen / Ineke Maas / Andrew Miles** (2002): *HISCO. Historical International Standard Classification of Occupations*, Leuven.

**Vladimir Iosifovič Levenštejn** (1966): "Binary Codes Capable of Correcting Deletions, Insertations, and Reversals". In: *Soviet Physics- Doklady* 10, 707–710. 129482234.

**Katrin Moeller** (2019): "Standards für die Geschichtswissenschaft! Zu differenzierten Funktionen von Normdaten, Standards und Klassifikationen für die Geisteswissenschaften am Beispiel von Berufsklassifikationen". In: *Aufklärungsforschung digital. Konzepte, Methoden, Perspektiven*. Hg. von Jana Kittelmann und Anne Purschwitz, Halle, 17-43.

**Katrin Moeller** (2021): "Ontologie historischer, deutschsprachiger Berufs- und Amtsbezeichnungen". In: *Websites des Historischen Datenzentrums Sachsen-Anhaltgeschichte.uni-halle.de*. 13.07.2021. [online].

**Heiko Müller / Johann-Christoph Freytag**: *Problems, Methods, and Challenges in Comprehensive Data Cleansing*. Berlin 2003. PDF. [online] 496492772.

**Wiebke Paulus / Britta Matthes** (2013): "Klassifikation der Berufe 2010 - Struktur, Codierung und Umsteigeschlüssel". *In: FDZ-Methodenreport*. Hg. von Forschungsdatenzentrum (FDZ) der Bundesagentur für Arbeit (BA) im Institut für Arbeitsmarkt- und Berufsforschung. Nürnberg. [online].

**Michael Piotrowski** (2012): "Natural Language Processing for Historical Texts". In: *Synthesis Lectures on Human Language Technologies* 5, H. 2, S. 1–157. 616519060

**Erhard Rahm / Hong Hai Do** (2000): "Data Cleaning: Problems and Current Approaches". In: *Bulletin of the Technical Committee on Data Engineering* 23, H. 4, S. 3–13. URN: urn:nbn:de:bsz:15-qucosa2-329680.