ZfdG

Zeitschrift für digitale Geisteswissenschaften

Beitrag aus:
Zeitschrift für digitale Geisteswissenschaften

Titel:
Automatisierte Identifikation und Lemmatisierung historischer Berufsbezeichnungen in deutschsprachigen Datenbeständen

Autor*in:
Jan Michael Goldberg

Kontakt: jan.goldberg@wiwi.uni-halle.de
Institution: Martin-Luther-Universität Halle Wittenberg, Lehrstuhl für empirische Makroökonomik
GND: 1240406630 ORCID: 0000-0002-4817-4283

Autor*in:
Katrin Moeller

Kontakt: katrin.moeller@geschichte.uni-halle.de
Institution: Martin-Luther-Universität Halle Wittenberg, Historisches Datenzentrum Sachsen-Anhalt, Institut für Geschichte
GND: 133366367 ORCID: 0000-0003-4090-5667

DOI des Artikels:
10.17175/2022_002_v2

Nachweis im OPAC der Herzog August Bibliothek:
1845604601

Erstveröffentlichung:
08.03.2022

Version 2.0:
20.07.2023

Letzte Überprüfung aller Verweise:
29.05.2023

Format:
PDF ohne Paginierung, Lesefassung

GND-Verschlagwortung:
Informations- und Dokumentationswissenschaft | Berufsforschung | Maschinelles Lernen | Automatische Klassifikation | Standardisierung |

Empfohlene Zitierweise:
Jan Michael Goldberg / Katrin Moeller: Automatisierte Identifikation und Lemmatisierung historischer Berufsbezeichnungen in deutschsprachigen Datenbeständen. In: Zeitschrift für digitale Geisteswissenschaften 7 (2022). 08.03.2022. Version 2.0 vom 20.07.2023. HTML / XML / PDF. DOI: 10.17175/2022_002_v2

Änderungen in Version 2.0 (20.07.2023):
Folgende Änderungen wurden vorgenommen: Sprachliche Verbesserungen im Text sowie inhaltliche Ergänzungen und Text und Bibliografie entlang der Monita der Gutachten.

Jan Michael Goldberg, Katrin Moeller

Automatisierte Identifikation und Lemmatisierung historischer Berufsbezeichnungen in deutschsprachigen Datenbeständen

Abstracts

Berufsangaben kommen in vielen historischen Quellen vor. Für eine Vielzahl von Forschungsgebieten ist nicht nur eine Standardisierung, sondern vor allem Klassifikation eine zentrale Voraussetzung zur Analyse. Dabei wird die Zuordnung von Schreibvarianten zu bereits definierten Gattungsnamen von Berufen in diesem Artikel als Lemmatisierung beziehungsweise Normierung bezeichnet, die Zuordnung der normalisierten Schreibweise zu einem Ordnungssystem als Klassifikation. Um hierbei manuellen Aufwand zu verringern, wird ein Algorithmus zur automatisierten Lemmatisierung historischer, deutschsprachiger Berufsangaben entwickelt. Das beste Ergebnis wird dabei mit einem Ansatz überwachten maschinellen Lernens erzielt. Insgesamt können etwa 72 Prozent der Berufsangaben lemmatisiert werden, etwa 98 Prozent dieser Zuordnungen sind korrekt.

Occupational information occurs in many historical sources. For a large number of research areas, not only standardization, but above all classification of these is a central prerequisite for analysis. In this article, the assignment of spelling variants to already defined generic names of occupations is referred to as lemmatization or normalisation, while the assignment of the normalised spelling and to a classification system is referred to as classification. In order to reduce manual effort, an algorithm for the automated lemmatization of historical, German-language occupational data is developed. The best result is achieved with a supervised machine learning approach. Overall, about 72 percent of the occupational data can be lemmatized, and about 98 percent of these assignments are correct.

1. Einleitung

Berufsangaben existieren in historischen Quellen an vielen Stellen und bilden eine wichtige Information über Menschen ab. Dadurch, dass es aber kein universelles System zur Erfassung von Berufen gab, existieren meist quellenspezifisch zusätzlich präfiguriert viele verschiedene Schreibweisen und Bezeichnungen nebeneinander. Gleichzeitig wurden in Gesellschaft und Wissenschaft verschiedene Ordnungssysteme gebildet, um über Klassifikationen von Berufen Informationen über ein Individuum zu strukturieren und zu ordnen. Die Auseinandersetzung mit Berufen ist auf vielen Ebenen wertvoll. In dieser Eigenschaft nimmt sie nicht nur für die Wirtschafts- und Sozialgeschichte, beispielsweise in Betrachtungen zur Entwicklung der Arbeit, eine besondere Relevanz ein. Die Klassifikation einzelner Berufsangaben stellt dabei eine fordernde Aufgabe dar, zumal kaum alle erdenklichen Schreibvarianten der Berufe manuell erfasst werden können. Eine automatisierte Zuordnung für historische deutschsprachige Berufe stellt eine hervorragende Lösung dar, um hier standardisierend für historische Quellen vergleichbare Ansätze zu bieten, auch wo die manuelle Klassifizierung als Aufwand zu groß erscheint.

Ziel dieser Abhandlung ist es darum, eine Möglichkeit aufzuzeigen, historische Berufsangaben automatisiert einem Klassifikationssystem zuzuordnen. Dabei wird die Zuordnung von Berufsangaben zu bereits definierten Varianten von Berufen hier als Lemmatisierung[1] bezeichnet. Bisher wird jedoch ein System zur automatisierten Lemmatisierung einer großen Menge historischer Berufsangaben vermisst. Entwickelt wird deshalb ein Algorithmus zur automatisierten Lemmatisierung dieser.

Besonders gehäuft kommen Berufsangaben in seriellen Quellen vor, die heute u. a. für genealogische Forschungen genutzt werden. Darunter fallen Kirchenbücher, Steuerregister, Adressverzeichnisse, Bürgeraufnahmeverzeichnisse oder verschiedene Amts- und Schöffenbücher. Die meisten der hier beschriebenen Berufsbezeichnungen stammen aus Quellen des 16. bis 19. Jahrhunderts und werden durch Berufsgattungsnamen der modernen Klassifikationssysteme ergänzt. Je älter diese Quellen sind, desto häufiger wird nicht unbedingt ein Beruf, sondern vielmehr ein Erwerbs- oder Berufsstand beschrieben. Historisch ist es einerseits von Interesse, diese Interpretation des ›Standes‹ als einen Definitionsansatz zu analysieren und zu ermitteln, welche unterschiedlichen Dimensionen ihn ausmachten. Dabei ist der Beruf nur eine Angabe unter anderen.[2] Diese Besonderheit der zeitbestimmten Definition, die von der heutigen Bestimmung des Berufes abweicht, macht es auch informationstechnisch zu

[1] Damit weicht der hier verwendete Begriff von der sprachwissenschaftlichen Terminologie ab, wo ein Lemma die Reduktion von Wortteilen auf die kleinste bedeutungstragende Einheit darstellt. Glück (Hg.) 2000, S. 403f.
[2] Moeller 2019, S. 23.

einer Herausforderung, die Bestandteile des Standes zu ordnen und sicher zu bestimmen. So finden sich in diesen Listen etwa für Frauen oder Kinder Informationen zum heutigen Familienstand (ledig, verheiratete, verwitwet, Sohn, Tochter etc.) für die Person oder in Relation zu einem berufsführenden Haushaltsvorstand.

Neben diesem Problem der historischen Ordnungssysteme lassen sich weitere informationelle Herausforderungen skizzieren, die bei der Lemmatisierung von originalsprachlichen historischen Begrifflichkeiten auftauchen. Die Verarbeitung ist insbesondere in genealogisch-prosopographischen Datenquellen aufgrund der hohen Dichte von Berufsangaben zeitaufwändig. Eine automatisierte Methode zur Umsetzung gibt es bisher vor allem für moderne (normierte) und englischsprachige Berufsangaben.[3]

Mit der Entwicklung einer Methode zur automatisierten Lemmatisierung von neuzeitlichen Standes- und Berufsangaben wird ein wichtiger Beitrag zu den Digital Humanities geleistet, weil mithilfe informatischer Lösungen die weitere Untersuchung historischer Fragestellungen unterstützt wird. Aufgrund der Besonderheiten, die den Angaben in jeder Sprache zuteilwerden, wird sich im Folgenden auf den deutschsprachigen Raum beschränkt. Als Klassifikationssystem wird eine bisher unveröffentlichte Beta-Fassung der Ontologie der historischen, deutschsprachigen Amts- und Berufsbezeichnungen (OhdAB)[4] benutzt, die auf der Methodik der Klassifikation der Berufe 2010 (KldB 2011)[5] basiert und diesen Ansatz um historische Berufsbezeichnungen erweitert. Dazu wird ein Algorithmus entwickelt, der für die weitere wissenschaftliche Arbeit in den verschiedensten Bereichen genutzt werden kann. Er stellt eine Methode dar, um zu einer Berufsangabe automatisiert Erkenntnisse über seine Klassifikation zu erhalten. Dabei wird der Algorithmus auf Berufsangaben in deutschsprachigen, neuzeitlichen, genealogisch-prosopographischen Quellen ausgelegt. Zur Entwicklung und Validierung werden Berufsangaben aus der Genealogischen Datenbasis (GEDBAS) genutzt. Jedoch können auch Berufsbezeichnungen anderer Quellen mit dem Algorithmus klassifiziert werden. Insbesondere bei großen Datenbeständen entfaltet ein automatisiertes Vorgehen erheblichen Nutzen. Bevor der Algorithmus vorgestellt wird, wird im nachfolgenden Abschnitt zuvor der Stand der Forschung beschrieben. Danach wird in seine technische Umsetzung eingeführt, bevor der Algorithmus validiert wird. Am Ende ist eine Zusammenfassung samt Ausblick zu finden.

2. Forschungsstand

Die Herausforderung eines Algorithmus zur automatisierten Kategorisierung von Berufsangaben besteht darin, sich unterscheidende Einträge, die die gleiche Sache beschreiben, zusammenzuführen. Bei dieser Aufgabe handelt es sich also im Wesentlichen um eine Dublettenerkennung, in der etymologisch identische, aber dennoch anders geschriebene Dubletten erkannt und zusammengeführt werden. Im ersten Unterabschnitt wird dazu einführend auf Berufsangaben im genealogischen Kontext eingegangen. Danach wird auf die Bereinigung und Lemmatisierung von Daten eingegangen, bevor abschließend die Besonderheiten der Berufsklassifikation in den Fokus gerückt werden.

2.1 Berufsangaben in genealogischen Quellen

Angaben zum Beruf und Stand waren in vielen historischen, personenbezogenen Quellen wie Kirchenbüchern obligatorisch. Diese Tendenz verstärkte sich mit der zunehmenden statistischen Erfassung des 19. Jahrhunderts, wobei erste Regularien entstanden, welche Standards für die Notation von Professionen entwickelten. Eine neue Etappe eröffnete sich mit der Säkularisierung des Personenstandswesens im Kaiserreich. So sah beispielsweise Preußen ab 1874 vor, »Stand oder Gewerbe« von Personen bei Geburt, Heirat und Todesfällen pflichtgemäß zu dokumentieren.[6] Mit der Entstehung eines um den Beruf herum organisierten Gesellschaftssystems im 19. Jahrhundert erhielt die Dokumentation von Stand und Gewerbe zentrale Funktionen für das Funktionieren des Staates,[7] das später auch von der Herausbildung von Institutionen zur Berufsklassifikation begleitet war. Zusätzlich konnte durch die Angabe des Berufs eine Unterscheidung zwischen namensgleichen Personen vorgenommen werden.[8] In der Folge ist es nicht verwunderlich, dass auch viele Genealogen diese Informationen erfassen. Neben den familiären Zusammenhängen und den Lebensdaten werden so auch Information zu Stand und Beruf den Datensätzen hinzugefügt.

---

[3] Cosca / Emmel 2010; Djumalieva et al. 2018; Gweon et al. 2017.
[4] Moeller et al. 2020. Die Klassifikation wurde bisher aufgrund von ausstehenden Qualitätsprüfungen noch nicht veröffentlicht, kann aber beim Historischen Datenzentrum Sachsen-Anhalt angefragt und genutzt werden.
[5] Bundesagentur für Arbeit (Hg.) 2021.
[6] Hinschius 1874, S. 41, 61f. u. 67.
[7] Kocka et al. 2000; Kohli 1985.
[8] Böhmen 1790, S. 29; Wurden Beruf oder Stand Jahrhunderte über in prosopographisch-genealogischen Quellen mitgeführt, wurde die Angabe von Berufen oder Titeln in Deutschland mit der Reformierung des Personenstandsrechts am 1. Januar 2009 abgeschafft, vgl. Schäfer 2006. Für künftige Forschungen entfällt damit eine wichtige Quelle.

Als Quasistandard zum Austausch solcher genealogischer Daten hat sich das GEDCOM-Format herausgebildet.[9] Eigenschaften von Personen werden in diesem textbasierten Format dazu mit sogenannten Tags versehen. Angaben zur Art der Arbeit oder des Berufs werden in dem Tag ›OCCU‹ zugeordnet.[10] Hier kann jedoch ein beliebiger freier Text eingetragen werden, sodass keine inhaltliche Prüfung über die Kompatibilität der Eintragung mit dieser Definition stattfindet.

2.2 Bereinigung und Lemmatisierung von Daten

Da Standesangaben also nicht zwingend nur Informationen zum Beruf enthalten – weder in den Primärquellen wie Kirchenbüchern noch in den aufbereiteten GEDCOM-Dateien –, ist eine Verarbeitung dieser Daten notwendig, um aus ihnen die relevanten Informationen zur Einordnung in ein berufliches Klassifikationssystem zu extrahieren. Zu diesem Zwecke wird folgend genauer auf die Datenbereinigung, Ähnlichkeits- und Distanzmaße sowie auf die Grundlagen von Klassifikationen eingegangen.

2.2.1 Datenbereinigung

Während der Datenbereinigung werden Fehler und Inkonsistenzen (im Folgenden auch ›Anomalien‹ genannt) erkannt und entfernt.[11] Beispielsweise können Rechtschreibfehler bestehen, Abkürzungen genutzt werden, Bezeichnungen in falsche Felder eingetragen werden oder eben zu viele Informationen darin vorhanden sein.[12] Fehler in Berufsangaben stellen in der Problemklassifikation nach Rahm und Do Einquellenprobleme (Quelle der Berufsangabe) auf einem Level einzelner Instanzen (Berufsangabe) dar. Wie oben bereits gezeigt, ist für historische Daten hier jedoch ebenso ein kontextualisierender Begriff des Berufsstandes wichtig. 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 sein kann.

Zur Lösung dieser qualitativen Probleme schlagen Müller und Freytag einen vierstufigen Prozess der Datenbereinigung vor.[13] 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*). Änderungen, die hier manuell vorgenommen werden, können in einem lernenden System jedoch einen bleibenden Effekt auf die Datenbereinigung haben. Insgesamt ist dieses Verfahren iterativ durchzuführen.

2.2.2 Ähnlichkeits- und Distanzmaße

Da insbesondere Berufsangaben in historischen Quellen verschiedene Schreibweisen aufweisen können, ist im Kontext der Anwendung eine Erkennung von Ähnlichkeiten zwischen diesen notwendig. Sollten zwei Bezeichnungen die gleiche Entität in der realen Welt repräsentieren, so sind sie Dubletten.[14] Da Berufsangaben Strings im Sinne einer semantischen Zeichenkette sind, 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 mögliche Verwandtschaft zwischen Wörtern aufzuzeigen.[15] Die Herausforderung, zwei Schreibvarianten desselben Wortes zu erkennen, ist ähnlich gelagert wie die Erkennung einer möglichen linguistischen Verwandtschaft zwischen zwei Wörtern. Da die Levenshtein-Distanz zudem die üblichste Methode zur Ähnlichkeitsanalyse zwischen zwei Strings darstellt,[16] wird sie auch im Weiteren Verwendung finden. Sie beschreibt die Anzahl von Löschungen, Einfügungen und Substituierungen einzelner Buchstaben, um von einem String zu einem anderen zu gelangen.[17]

---

[9] Gellatly 2015; Harviainen / Björk 2018, S. 4.
[10] Church of Jesus Christ of Latter-day Saints 2019, S. 90.
[11] Rahm / Do 2000, S. 1.
[12] Rahm / Do 2000, S. 3f.
[13] Müller / Freytag 2003, S. 10–13.
[14] Krause 2012, S. 14f.
[15] Dunn 2015, S. 196.
[16] Piotrowski 2012, S. 71.
[17] Levenštejn 1966.

Daneben gibt es auch andere Ähnlichkeitsmaße, deren Verwendung auf ähnliche Aufgabenstellungen sinnvoll erscheinen könnte. Beispiele dafür sind die Jaro-Winkler-Distanz, eine stochastisch gewichtete Levenshtein-Distanz[18] oder Deep-Learning-Algorithmen wie DeezyMatch und STANCE.[19] Zudem können phonetische Ähnlichkeitsmaße wie die Kölner Phonetik eingesetzt werden. Weitere Möglichkeiten sind die Heuristiken von Bryan Jurish für DTA::CAB.[20] Auch Machine-Learning-Applikationen wie bei Bollmann und Domingo / Casacuberta können Einsatz finden. In dieser Arbeit kann keine Aussage darüber getroffen werden, welche Methode in welchem Fall die besten Ergebnisse produziert. Im Zweifel kann die Ähnlichkeitsanalyse im Algorithmus und Programmcode verändert werden. Bei einer Veränderung des Programmcodes ist es wichtig, auch eine Anpassung der Grenzwerte vorzunehmen, wie im weiteren Verlauf des Textes deutlich wird.

2.2.3 Grundlagen von Klassifikationen

Unter der Klassifikation wird die Gliederung von Elementen einer Menge in verschiedene Klassen nach einer definierten Logik verstanden. Dieser Vorgang kann automatisiert werden, wenn die zugrundeliegenden Gesetzmäßigkeiten in einem Algorithmus Formalisierung finden. Eine (automatisierte) Klassifikation kann dabei entweder ein Objekt einer Klasse zuordnen oder eben auch dahingehend scheitern, dass keine Klasse ausgewählt werden kann. Die Zuordnung zu (k)einer Klasse kann zudem korrekt oder nicht korrekt sein. Durch diese binäre Ansicht ergeben sich vier mögliche Kombinationen (siehe Tabelle 1). Als erstrebenswert gilt dabei eine Erhöhung der TP- und TN-Ergebnisse. FP- und FN-Ergebnisse sind hingegen zu vermeiden.

	Klassifikation korrekt	Klassifikation nicht korrekt
Klassifikation erfolgt	True positive (TP)	False positive (FP)
Klassifikation nicht erfolgt	True negative (TN)	False negative (FN)

Tab. 1: Konfusionsmatrix zur Klassifikation in Anlehnung an Fawcett 2006. [Goldberg / Moeller 2022]

Durch die Kombination der Anzahl der jeweiligen Zustände kann die Güte der Klassifikation bewertet werden. Dies ist notwendig, weil ein hoher Anteil von Treffern oftmals auch mit vielen falschen Ergebnissen einhergeht – bei keinen Treffern hingegen kann auch kein Treffer falsch sein. Eine Möglichkeit zur Ermittlung der Qualität einer Klassifikation stellt das F1-Maß dar.[21] Dieses wird genutzt, um ein optimiertes Verhältnis zwischen den gefundenen Treffern und den richtigen Treffern zu erzielen. Hierbei werden die Trefferquote (R, für *recall*) und die Genauigkeit (P, für *precision*) der Klassifikation gemäß der Formel für das F1-Maß (siehe Formel 1) in eine Beziehung gesetzt.

Formel 1: $F1=2*\frac{P*R}{P+R}$

Sind hier die Genauigkeit und die Trefferquote beim F1-Maß gleich gewichtet, so ist auch jede andere Gewichtung denkbar. Die Genauigkeit ergibt sich aus Formel 2, die Trefferquote aus Formel 3.

Formel 2: $P=\frac{|TP|}{|TP|+|FP|}$ Formel 3: $R=\frac{|TP|}{|TP|+|FN|}$

Anders als bei einer manuellen Klassifikation, bei der die Korrektheit einer Zuordnung vorher ermittelt wird, ist das bei einer automatisch durchgeführten Klassifikation – wenn überhaupt – erst im Nachgang möglich. Jedoch verändert jede zusätzliche Schreibvariante, die einem Lemma zugeordnet wird, die Eigenschaften dieser Klasse. Dadurch, dass künftige Klassifikationen auf diese vorherigen Informationen zugreifen können, findet überwachtes Lernen statt.

2.3 Berufsklassifikationen

Grundsätzlich muss zwischen der Lemmatisierung von Berufsbezeichnungen und der Klassifikation von Berufen unterschieden werden. Mit Ersterer, der Lemmatisierung der Bezeichnungen zu Berufen, befasst sich dieser Artikel. Dabei wird eine Vielzahl von Schreibvarianten einem normierten Berufsnamen zugeordnet, sofern eine bestimmte sprachliche Übereinstimmung erkennbar ist. Diese Berufsnamen können in einem weiteren, übergeordneten Klassifikationssystem auch inhaltlich-analytisch

[18] Vgl. Hauser / Schulz 2007.
[19] Vgl. Hosseini et al. 2020; Tam et al. 2019.
[20] Vgl. Jurish 2012.
[21] Christen / Goiser 2007, S. 140f.

zu verschiedenen Berufsgruppen geordnet werden, indem das Definitionskriterium der Tätigkeit zur Klassifikation herangezogen wird. In solche Systeme wird in diesem Abschnitt eingeführt. Relevant ist das übergeordnete System der Berufsklassifizierung, weil es die Entitäten determiniert, auf denen die nachfolgende Entwicklung des Algorithmus aufbaut.

Zur Klassifikation von Berufen existieren verschiedene Ansätze die bisher vor allem moderne internationale,[22] moderne deutschsprachige[23] oder historische englischsprachige[24] Berufsnamen führen. Von diesen Standards wird häufig eine Vielzahl forschungsbasierter Klassifkationsansätze für unterschiedliche Analysen abgeleitet. In Hinblick auf die Entwicklung von Datenstandards nach FAIR-Prinzipien werden solche kompatiblen Systeme zukünftig höheres Gewicht besitzen, weil die Anbindung an Standards die Nachvollziehbarkeit und Vergleichbarkeit von Forschungsergebnissen gewährleistet.[25] Im deutschsprachigen Raum ist vor allem die Klassifikation der Berufe 2010 (KldB 2010) beziehungsweise jetzt 2020 zu nennen. Die Methodik der KldB 2010 wurde von Katrin Moeller auf viele historische, deutschsprachige Berufsbezeichnung angewendet.[26] Dieses System wird im Weiteren Anwendung finden, da es für den deutschsprachigen Raum die umfangreichste Lösung darstellt. Für diese Arbeit wurde der Stand der OhdAB vom 27. Mai 2020 verwendet (mit 183.381 Varianten). Alternativ dazu könnte HISCO in Betracht gezogen werden. HISCO stellt die historische Erweiterung von ISCO 68 dar. Davon wird an dieser Stelle abgesehen, weil auf der offiziellen HISCO-Webpräsenz derzeit nur 1.306 deutsche Berufsbezeichnungen genutzt werden, während die OhdAB momentan 44.893 Normbezeichnungen für deutschsprachige Berufe führt.[27] Zudem enthält HISCO keine umfangreiche Zuordnung von Varianten eines Berufs, wodurch die Zuordnung zu historischen Berufsangaben erschwert wird. Damit bleiben viele Berufsnamen bisher ohne sichere Zuordnung in der HISCO. Durch die Granularität der KldB 2010 kann zwar jeder Beruf der HISCO in der KldB abgebildet werden, nicht jedoch andersherum. Des Weiteren existieren im deutschsprachigen Raum historische Berufsklassifikationen,[28] die in die OhdAB mit eingeflossen sind. Beachtlich ist zudem die Systematisierung des Thesaurus Professionum von 23.000 Berufen, die auf Erschließungen von Leichenpredigten der Forschungsstelle für Personalschriften der Philipps-Universität Marburg zurückgehen.[29]

2.3.1 Klassifikation der Berufe 2010

Die KldB 2010 teilt Berufe nach einer fünfgliedrigen Hierarchiestruktur ein.[30] Der Einsteller (Berufsbereiche) gliedert die Berufe in grundlegende Themen.[31] Die nächsten drei Ebenen (Berufshauptgruppen, Berufsgruppen und Berufsuntergruppen) beschreiben die berufsfachlichen Zusammenhänge.[32] Je stärker zusammenhängende Fähigkeiten, Tätigkeiten und Kompetenzen zwischen Berufen existieren, desto näher sind sich diese in der Hierarchie. Zuletzt beschreibt der Fünfsteller (Berufsgattungen) das Anforderungsniveau, sodass durch ihn unterschiedliche Komplexitätsgrade desselben Berufs ausgedrückt werden können.[33] Insgesamt existieren auf der Ebene des Fünfstellers mittlerweile 1.900 Berufsgattungen.[34]

[22] International Standard Classification of Occupations (ISCO), ILO (Hg.) 2021.
[23] Klassifikation der Berufe (KldB), Bundesagentur für Arbeit (Hg.) 2021.
[24] Historical International Standard Classification of Occupations (HISCO); van Leeuwen et al. 2002.
[25] Moeller 2019.
[26] Moeller et al. 2020.
[27] International Institute of Social History (Hg.) 2020.
[28] Vgl. Schüren 1989; Arbeitskreis für Wirtschafts- und Sozialgeschichte Schleswig-Holsteins 1991.
[29] Philipps-Universität Marburg, Forschungsstelle für Personalschriften (Hg.) 2021.
[30] Bundesagentur für Arbeit (Hg.) 2011, S. 16.
[31] Paulus / Matthes 2013, S. 7.
[32] Paulus / Matthes 2013, S. 8.
[33] Paulus / Matthes 2013, S. 9f.
[34] Bundesagentur für Arbeit (Hg.) 2011, S. 18.

Stellensystem	Bezeichnung für das Beispiel des Bäckers	Gruppenbezeichnung	Anzahl der Gruppen über alle Berufsgattungen
1-Steller B 29222	Rohstoffgewinnung, Produktion und Fertigung	Berufsbereiche	10 Gruppen
2-Steller B 29222	Lebensmittelherstellung und -verarbeitung	Berufshauptgruppen	72 Gruppen
3-Steller B 29222	Lebensmittel- und Genussmittelherstellung	Berufsgruppen	260 Gruppen
4-Steller B 29222	Berufe in der Back- und Konditoreiwarenherstellung	Berufsuntergruppen	941 Gruppen
5-Steller B 29222	Berufe Back- und Konditoreiwarenherstellung - fachliche Tätigkeit	Anforderungsniveau	1.900 Gruppen

Tab. 2: Nummernsystem der KldB 2010 / OhdAB am Beispiel des Berufes Bäcker. [Goldberg / Moeller 2022]

Einzelne Berufe sind in der KldB 2010 nicht aufgeführt, sondern in die entsprechenden Berufsgattungen einzuordnen; dennoch bietet diese Lösung bereits eine gute Näherung an moderne Individualbezeichnungen.

2.3.2 Erweiterung um historische Berufe

Die Methodik der KldB 2010 ist grundsätzlich auch auf historische Berufe anwendbar, weil sie nach Tätigkeiten und Anforderungsniveaus ordnet, die auch für vergangene Arbeitsfelder erschließbar sind. Mit der OhdAB liegt eine solche Grundlage zur Klassifikation von historischen Berufs- und Amtsbezeichnungen in einer Beta-Fassung vor. Dabei werden alle Schreibvarianten (unter Vergabe einer fortlaufenden ID) von Standesbezeichnungen nach der Methode der KldB 2010 erfasst und zu einem Berufsgattungsnamen (Zusatz einer dreistelligen Individualnummer) sowie einer fünfstelligen Klassifikation (Klassifikationsnummer) angeordnet.

Der ursprünglichen Fassung der KldB 2010 wurden dem Berufsgattungsnamen fortlaufend die Klassifikationsnummern unter einem Wert von 500 zugewiesen, historische Berufe erhielten bei der Ergänzung einen Wert größer als 500, wodurch die modernen und historischen Gattungsnamen voneinander differenzierbar bleiben. Die KldB 2010 wurde zudem um einige wenige Berufsgruppen ergänzt, die sich in das Konzept der ursprünglichen Fassung nicht einfügen ließen. Dies gilt etwa für die Gruppe von Stadt- und Hofwachen, die weder dem Personen- und Objektschutz, der Polizei noch dem Militär zugeordnet werden konnten. Gleiches gilt für die Hofverwaltung, militärische Berufsgruppen oder das Landhandwerk. Insgesamt folgt die Klassifikation jedoch der Methodik der KldB 2010. Zudem wurden allgemeinere Beschreibungsgruppen (wieder-)eingeführt, um auch Gattungsbegriffe wie ›Beamter‹ oder ›Arbeiter‹ einzuordnen. Dies ist aufgrund des spezifischen Tätigkeitskonzeptes der KldB 2010 ansonsten nicht möglich. Dieser Kennung vorangestellt wird ein A oder B. Der weitaus geringere Teil ist mit A betitelt (bisher ca. 600 Gattungsbegriffe), wodurch solche Angaben kenntlich gemacht werden, die in den historischen Registern eine Eintragung zum Stand verfügen, der heute aber keinen Beruf mehr definiert. Wie oben beschrieben waren dies in der Regel Verwandtschaftsverhältnisse zu einem Haushaltsvorstand. Sehr häufig handelt es sich um Angaben zur Kenntlichmachung der Armut einer Person oder zum Bezug von Almosen, Altenteil, Renten- oder Invalidenbezügen. Gleichzeitig kommen Angaben zu Eigentums- und Besitzverhältnissen, Religion, Rechts- und Einwohnerbezeichnungen vor. Ein B weist demnach darauf hin, dass es sich um einen Beruf im Sinne eines Tätigkeitskonzepts handelt. Insgesamt sind fast 44.582 normierte Berufsschreibweisen so klassifiziert.

Die Liste der Varianten hingegen besteht aus möglichen Schreibvarianten der Berufe, die einer Normschreibweise eines Berufs der Konkordanz (Auflistung aller möglichen Berufe als Normschreibweise) zugeordnet ist. Es besteht eine 1:n-Beziehung, da ein Eintrag der Konkordanz beliebig viele Varianten aufweisen kann. Unterschiede zwischen Normschreibweise und Variante lassen sich an verschiedenen Aspekten erkennen. So enthält die Normschreibweise eine geschlechtsübergreifende Schreibweise (z. B. ›Müller/in‹), die Varianten allerdings die Berufe je Geschlecht einzeln separiert (hier ›Müller‹ und ›Müllerin‹). Insgesamt sind derzeit weit über 300.000 Varianten erfasst. Die Liste der Varianten wird durch das Historische Datenzentrum Sachsen-Anhalt jedoch stetig erweitert. Für die weitere Arbeit wird ein Auszug aus diesen Varianten verwendet, der zur Validierung näher beschrieben wird.

3. Entwicklung des Algorithmus

Die Entwicklung eines Algorithmus ist notwendig, um die Vorgehensweisen hinter der Lemmatisierung der Berufsangaben – und somit die zugrundeliegenden Heuristiken – formalisiert zum Ausdruck zu bringen. Dazu werden zunächst die Anforderungen an diese Automatik detaillierter beleuchtet. Danach folgt eine Umsetzung der Schritte der Datenbereinigung nach Müller und Freytag.[35]

3.1 Anforderungen das Ergebnis

Zunächst sollen möglichst viele Berufsangaben den richtigen Entitäten, im Weiteren ›Klassen‹, zugeordnet werden. Ein Beruf 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.

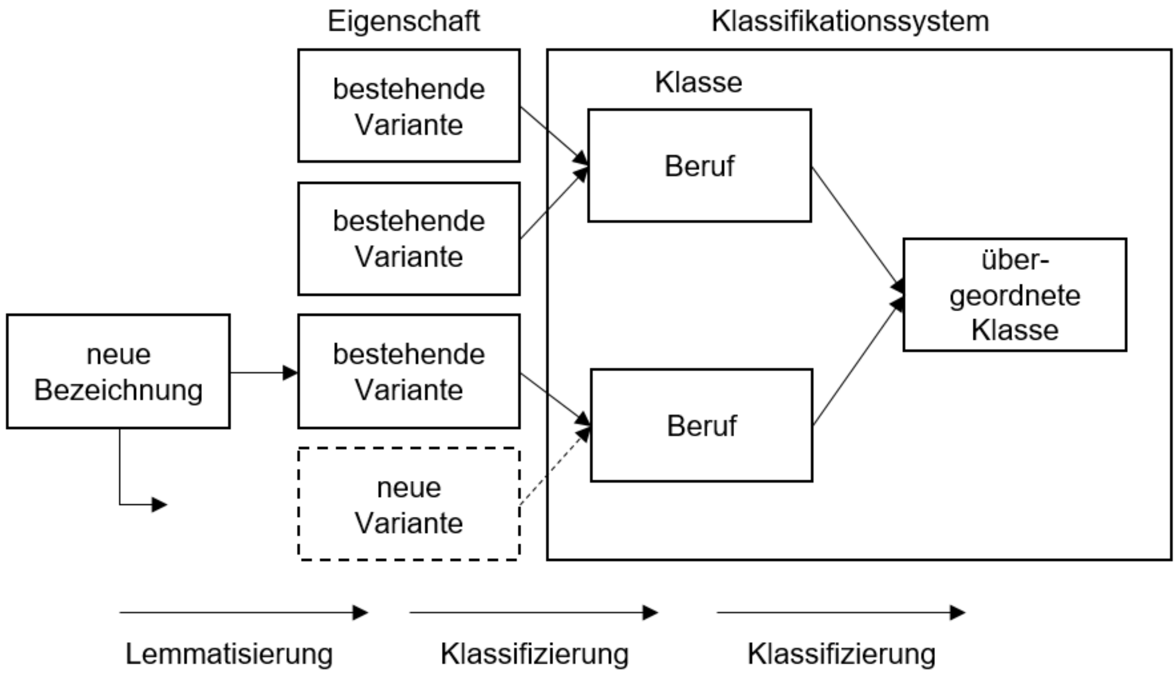

Abb. 1: Begriffe und Zusammenhänge des Algorithmus. [Goldberg / Moeller 2022]

Eine Erhöhung der TP-klassifizierten (neuen Bezeichnungen) allein geht jedoch oftmals auch mit der Erhöhung von FP-Klassifizierungen einher. Aus diesem Grund wird nicht die Anzahl der TP-Klassifizierungen optimiert, sondern das F1-Maß. Zudem soll die Klassifizierung automatisch geschehen, eine manuelle Überprüfung des Ergebnisses geschieht infolgedessen nicht. Das ist notwendig, um große Datenbestände mit hunderttausenden Berufsbezeichnungen 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 soll in einem Programmcode umgesetzt werden, der in weiteren Applikationen eingebunden werden können soll.

Der Algorithmus soll zwar mit Hilfe genealogisch-prosopographischer Quellen validiert werden, jedoch danach auch auf andere Berufsangaben angewendet werden können. Art und Umfang der Quelle sind dabei nicht entscheidend. Wichtiger ist es, dass es sich um deutschsprachige Berufsangaben aus dem Zeitraum der Neuzeit (ab ca. 1500) handelt. Bei anderen Angaben steigt die Wahrscheinlichkeit, dass der Algorithmus keine verwertbaren Ergebnisse liefert (z. B. bei lateinischen Angaben), jedoch soll eine nachträgliche Erweiterung der Sprachen möglich sein.

[35] Müller / Freytag 2003.

Des Weiteren können Datenfelder zum Beruf mit verschiedenen Informationen gefüllt sein. In vielen Fällen dürften sie als Freitextfeld keiner Konsistenzprüfung unterzogen worden sein. Das führt dazu, dass prinzipiell alles in einem solchen Feld stehen kann. Es ist eine Anforderung, daraus den Beruf zu separieren. Möglicherweise sind auch mehrere Berufsangaben verzeichnet, die dann getrennt voneinander erkannt werden sollten. Auch berufsferne oder berufsfremde Informationen in den Berufsangaben sollen jeweils separiert werden (z. B. der fälschlicherweise in einem Datenfeld für die Berufsangabe angegebene Wohnort). Lemmatisiert wird jedoch nur die Angabe zum Stand und / oder Beruf. Mögliche berufsferne, separierte Informationen unterliegen keiner weiteren Interpretation.

3.2 Methodik der Datenbereinigung

Wie oben beschrieben, setzt sich die Datenbereinigung aus verschiedenen Schritten zusammen, die nun nacheinander durchgeführt werden. Zunächst wird im data auditing der zu bearbeitende Datensatz betrachtet. Die hier entwickelte Datenbereinigung soll allgemein auf deutschsprachige Berufsbezeichnungen anwendbar sein. Dazu werden Berufsangaben genutzt, die in öffentlich zugänglichen GEDCOM-Dateien gespeichert sind. Im deutschsprachigen Raum stellt GEDBAS eine der wesentlichen Sammlungen von genealogischen Daten dar. In dieser Datenbank sind in etwa 13.000 Dateien und 22 Millionen Personen abgebildet.[36] Ein Teil dieser Dateien ist von den Autoren zum öffentlichen Download freigegeben. Die Ausführung eines Scrapers zur Sammlung der öffentlichen GEDCOM-Dateien in GEDBAS am 14.04.2020 erbrachte 2.899 Dateien.[37] Um die Berufsangaben aus den Dateien zu erfassen, werden aus allen GEDCOM-Dateien die Berufsangaben (›OCCU‹-Tag) ausgelesen und in einer Liste zusammengefasst.[38] Insgesamt werden auf diese Weise 229.669 Berufsangaben ermittelt. Nach einer Eliminierung der mehrfachen Angaben bleiben 60.000 verschiedene Bezeichnungen übrig. Dabei werden doppelte Token gelöscht, sodass jeder Type einer Bezeichnung in der Liste nur einmal vorkommt.

Weitere naheliegende, aber nicht in den GEDCOM-Daten auftretende Anomalien bei Berufsangaben werden ebenso mit eingebunden. Das ist darin begründet, dass der Algorithmus auf alle deutschsprachigen Berufsangaben seit ca. 1500 anwendbar sein soll und ggf. mögliche Anomalien in den GEDCOM-Daten strukturell komplett fehlen. Die folgenden Anomalien wurden insgesamt erkannt:

- **Mehrere Berufe**: In einer Berufsangabe kann ein Verweis auf mehrere Berufe vorhanden sein, beispielsweise, weil die Bezugsperson verschiedene Berufe in ihrem Leben (hintereinander oder parallel zueinander) ausgeübt hat. Verschiedene Berufsangaben können durch Trennoperatoren abgegrenzt sein. Eine Besonderheit ergibt sich bei der Verwendung von Ergänzungsstrichen in einer Berufsangabe (z. B. ›Gold- und Silberschmied‹).
- **Abkürzungen**: Berufe können eine Abkürzung erfahren, die wiederum sehr quellenspezifisch sein können. So ist es denkbar, dass ein ›B.‹ für ›Bürger‹ steht, aber auch für ›Bauer‹. Abkürzungen enden in vielen Fällen, allerdings nicht immer, mit einem Punkt.
- **Rollen**: Dem Beruf vor- und nachgestellt können weitere Angaben zur Rolle innerhalb des Berufsbildes sein. Das betrifft im Handwerk beispielsweise die fünf Qualifikationsstufen von Hilfsarbeitern, Burschen und Knechten, Lehrlingen und Gesellen, Altgesellen und arbeitenden Fachhandwerkern ohne Meistertitel, Meistern sowie Obermeistern beziehungsweise Oberältesten von Innungen.
- **Zeitangaben**: Zur Spezifizierung des Zeitpunktes der Bezeichnung mit einem Beruf kann eine Zeitangabe aufgenommen werden. Diese ist möglicherweise durch Klammern abgegrenzt. Auch die Verwendung von Ziffern ist ein Indikator für eine Zeitangabe. Jedoch können Ziffern regulärer Bestandteil der Berufsbezeichnung sein (z. B. ›1. Pfarrer‹ oder ›2. Offizier‹). Daneben könnten temporale Präpositionen auf Zeitangaben hinweisen. Bei der Angabe von konkreten Daten oder Jahreszahlen kommen temporale Präpositionen nach dem Beruf (z. B. ›Bauer im Jahre 1873‹) wie auch zu Beginn (z. B. ›am 02.03.1734: Hufschmied‹) vor. Häufig stehen Zeitangaben auch ohne Präposition.
- **Berufsstatus**: Gleichfalls können temporale Informationen darüber vorhanden sein, ob der benannte Beruf aktiv ausgeübt wird oder es sich um einen vormaligen Beruf handelt. So existieren Möglichkeiten, den Status einer Person in Bezug auf den Beruf zu beschreiben (z. B. ›pensioniert Lehrer‹ oder ›*g*ewesener Gerichtsschreiber‹). Hinzu treten Bezeichnungen wie ›Altenteiler‹ oder ›Invalid‹, die aber keine spezifischen Berufsangaben mehr enthalten.
- **Quellenangaben**: Analog zur Angabe eines Zeitpunkts ist auch der Verweis auf Quellen möglich. Quellen können auf verschiedene Arten angegeben werden. Ein vorkommender Fall ist die Verwendung von URLs oder HTML-Codes für Hyperlinks, um auf Inhalte im Internet zu verweisen.
- **Ortsangaben**: Häufig kommt auch die Angabe des Ortes einer Berufsausübung vor. Anders als bei Zeitangaben werden bei den Ortsangaben (lokale) Präpositionen wie ›in‹, ›bei‹ oder ›von‹ häufig verwendet. Neben dem Arbeitsort kann es auch

[36] GEDBAS, Verein für Computergenealogie (Hg.) 2021.
[37] Siehe den entsprechenden Programmcode in (Online-Repo). [verlinken]
[38] Siehe den entsprechenden Programmcode in (Online-Repo). [verlinken]

vorkommen, dass der Herkunfts- oder Wohnort genannt wird, der ebenfalls nicht zu Beschreibung der Tätigkeit genutzt werden kann.

- **Arbeitgeber**: Ebenfalls kann der Arbeitgeber genannt werden. Die Präpositionen ähneln dabei denen der Ortsangaben (z. B. ›Kalkulator bei der Deutschen Versicherung A.G.‹). Darunter ist auch die Zuordnung zu einem Dienst- oder Lehnsherren oder einem Regiment etc. zu verstehen. Bei Berufsangaben werden in diesem Sinne auch Zusätze wie ›herrschaftlich‹ oder ›königlich‹ als solche betrachtet. Im Militär dominieren hier Angaben zu Regimentern etc.
- **Familienstand**: Ein Datenfeld, welches mit ›Stand und Gewerbe‹ überschrieben ist, lässt vielerlei Möglichkeiten zu. Eine davon ist der Familienstand. Hierrunter fällt im engeren Sinne, ob eine Person ledig, verheiratete oder verwitwet ist. Bezeichnungen für unverheiratete Frauen sind so beispielsweise ›Jungfer‹ oder ›Jungfrau‹, bei Männern dahingegen ›Junggeselle‹ oder ›Geselle‹. Manche Angaben können auch darauf hinweisen, dass die Berufsangabe sich nicht direkt auf den Stelleninhaber bezieht, sondern auf eine nahestehende Person. So kann die Rolle in der Familie benannt sein (z. B. ›Sohn‹ oder ›Tochter‹). Die die Bezeichnung als ›Witwer‹ oder ›Witwe‹ ist erwähnenswert.
- **Rechtsstatus**: Der Rechtsstatus einer Person kann ebenso Teil einer personenstandlichen Aussage sind. Eine wesentliche, oft vorkommende Unterscheidung hierbei ist die zwischen ›Bürgern‹ und ›Inwohnern‹ oder ›Einwohnern‹.
- **Besitzinformationen**: Auch kann die Angabe Informationen über den Besitz des Beschriebenen enthalten, ohne dass aus diesen direkt (ohne weitere Annahmen) ein Rückschluss auf die berufliche Tätigkeit gezogen werden kann (z. B. ›Hausbesitzer‹ oder ›Fabrikbesitzer‹, ›Erbe‹).
- **Titelangaben**: Vom Beruf (und auch dem Rechtsstatus) abzugrenzen sind Titel wie Adelstitel oder akademische Titel. Am häufigsten kommt hierbei die vorangestellte Angabe des Doktorgrades vor. Auch können weitere Adjektive wie ›wohlgeachtete‹ oder ›ehrbare‹ vorangestellt werden oder auch ehrende Anreden (›Herr‹ / ›Frau‹) enthalten.
- **Fremdsprachliche Angaben**: Auch fremdsprachliche Angaben können vorkommen. Aufgrund des derzeitigen Fokus der OhdAB auf historische deutschsprachige Berufe und der speziellen Konzentration auf deutschsprachige Quellen ist die entwickelte Lösung nicht besonders geeignet für Berufs- und Standesbezeichnung anderer Sprachen. Sie werden nicht mit lemmatisiert und klassifiziert und deshalb als TN-Ergebnisse erkannt. Langfristig ist es ein erstrebenswertes Ziel, eine Mehrsprachigkeit (besonders lateinische Berufsbezeichnungen) zu implementieren. In den GEDCOM-Daten kommen insbesondere niederländischer Bezeichnungen häufig vor. Aufgrund der sprachlichen Nähe zum Deutschen stellen diese eine besondere Herausforderung dar.
- **Tippfehler und Schreibvarianten**: Besonders nachfolgende (oder führende) Leerzeichen kommen häufig vor, weil sie bei der Dateneingabe schnell übersehen werden können, dennoch aber Teil des Strings sind. Denkbar sind auch sonstige Tippfehler, fehlende, zusätzlich vorhandene oder vertauschte Zeichen. Grundsätzlich wird von Schreibvarianten gesprochen, ohne die Herkunft dieser (quellenbasiert, transkriptionsbasiert) für bestehende Sammlungen aufgrund fehlender direkter Bezugsebenen zwischen Quellen und Datensammlung nicht überprüfbar sind. Eine besondere Form von Schreibvarianten stellen Durchkoppelungen dar. Sie können überall dort vorkommen, wo verschiedene Morpheme aneinandergesetzt werden, was bei Berufsangaben vergleichsweise häufig der Fall ist. Beispiele dafür sind der ›Reserveoffizier-Anwärter‹, ›Bäcker-Meister‹ oder ›Gerichts-Gehilfe‹. In seltenen Fällen wird der Bindestrich auch als Trennungsoperator zwischen verschiedenen Berufen genutzt (z. B. ›Häusler-Weber‹).
- **Falsche Verwendung des Feldes**: Inhaltlich falschen Angaben, die mit einer Berufsangabe nichts zu tun haben, kann die falsche Verwendung des Datenfeldes zugrundeliegen. Wahrscheinlicher als eine bewusste Fehlinterpretation ist vermutlich die versehentliche Vertauschung, u. a. mit Datenfeldern für Namen, Wohnorte oder Datumsangaben.

Der Umgang mit diesen wird nachfolgend in der workflow spezification festgelegt. Dabei handelt es sich um die Formalisierung von Heuristiken zur Interpretation der Anomalien. Die Spezifizierung des Ablaufs der Datenbereinigung wird in drei Teile gegliedert: Zunächst findet (1.) eine grundsätzliche Vorverarbeitung der ursprünglichen Berufsangabe statt. Danach werden (2.) verschiedene, darin enthaltene Berufsangaben voneinander separiert. Abschließend erfahren diese Strings (3.) eine weitere Nachbearbeitung, indem berufsfremde Angaben separiert werden. Die Reihenfolge der einzelnen Schritte ist relevant und zu beachten. Nachfolgende Schritte können zu anderen Ergebnissen führen, sollten die vorhergehenden nicht zuvor ausgeführt worden sein.

3.3 Ablauf der Datenbereinigung

3.3.1 Normieren von Trennoperatoren

Es gibt verschiedene Operatoren, die voneinander abzugrenzende Informationen innerhalb der Berufsangabe trennen. Mögliche Trennoperatoren sind:

- u.
- +

- ,
- ;
- &
- /
- -

Dabei können verschiedene Kombinationen mit vor- oder nachgestellten Leerzeichen Aufschluss über den spezifischen Zweck des Zeichens geben. Beispielsweise stellt der Bindestich nur ohne vorangehendes Leerzeichen, einen Trennoperator da, da er ansonsten als Ergänzungsstrich interpretiert werden sollte. Falls hinter dem Ergänzungsstrich zusätzlich statt einem Leerzeichen ein Komma gesetzt ist, handelt es zudem um eine Aufzählung, was in der Ermittlung des entsprechenden Wortteils zu beachten ist. Ausgenommen von der Trennung ist die Kombination ›- und‹ wie beispielsweise in ›Gold- und Silberschmied‹. Hier wird von einem zusammenhängenden Begriff ausgegangen.

Es ist für die nachfolgende Verarbeitung hilfreich, wenn diese Operatoren normiert und durch einen einzigen Trennoperator getrennt werden. Die Trennoperatoren werden durch ein ›und‹ ersetzt. An den Stellen, an denen infolgedessen ein ›und‹ steht, erfolgt eine Trennung des Strings unter der Löschung von des vormals verbindenden ›und‹. Jeder der entstehenden Teile wird datentechnisch abgegrenzt, sodass dieser zwar einzeln behandelt werden kann, dennoch aber auch die ursprüngliche Zusammengehörigkeit nachvollziehbar bleibt. Das ist aus dem Grunde sinnvoll, da in den dann getrennten Teilen neben dem Beruf weitere (berufsferne) Informationen stehen könnten. Diese sind für eine unmittelbare Klassifikation des Berufs nicht notwendig (oder gar hinderlich), sollen aber im Bezug zur Berufsangabe dennoch nicht verloren gehen, da sie ggf. wichtige weiterführende Informationen erhalten. Die Informationen werden für eine anschließende Interpretation separiert und damit von der eigentlichen Berufsangabe getrennt. Dennoch bleibt auch hier die Zusammengehörigkeit nachvollziehbar.[39]

Es folgen Schritte zur Separierung berufsferner Angaben aus der Bezeichnung. Teilweise wird die berufsferne Angabe durch die Separierung aus der eigentlichen Berufsangabe gelöscht, mitunter aber auch beibehalten, weil sie für die Lemmatisierung von Relevanz ist. Separiert wird in zwölf Kategorien:

- Beruf
- Rolle
- Jahr
- URL
- Ort
- Arbeitgeber
- Familienstand
- Rechtsstatus
- Besitzinformation
- Titel
- Berufsstatus
- Weiteres

3.3.2 Entfernung von Leerzeichen

Leerzeichen, die am Anfang oder am Ende des Strings stehen, werden entfernt.

3.3.3 Auflösung von Abkürzungen

Die Abkürzungen können je nach Quelldaten sehr unterschiedlich gewählt worden sein. Es ist empfohlen, oft vorkommende und konsistent verwendete Abkürzungen in der Quelle im Programm zu ergänzen. Einige Abkürzungen, die in den GEDCOM-Daten vorkommen und allgemeingültig erscheinen, werden an dieser Stelle dennoch bereits aufgenommen. Kommen sie vor, werden sie aufgelöst. Das bedeutet, dass dieses ausgeschrieben werden. Dies sind:

- ›Bgmst.‹ für ›Bürgermeister‹
- ›Ing.‹ für ›Ingenieur‹

[39] Beispielsweise wird die Bezeichnung ›Hutmacher und Bürger‹ in zwei Teile separiert, wobei der Bürger keine Berufsbezeichnung darstellt. Für eine mögliche nachfolgende Analyse ist es ggf. von Relevanz, nachzuvollziehen, dass der Hutmacher einen Bürgerstatus innehatte.

Nicht möglich ist eine solch allgemeine Übersetzung von Abkürzungen beispielsweise bei der Angabe ›B.‹, die mit einer großen Wahrscheinlichkeit für ›Bauer‹ oder ›Bürger‹ stehen könnte. Die Verwendung aller Abkürzungen aus den GEDCOM-Daten würde zu einem Overfitting führen. Die Abkürzung ›Dr.‹ dahingegen wird – trotz eindeutiger Verwendung – bewusst so belassen, da sie später als Titelangabe separiert wird. Auch trifft dieses auf die Abkürzungen ›a. D.‹ und ›i. R.‹ zu, da diese den Berufsstatus beschreiben. Ebenso werden weitere Abkürzungen, die für bestimmte Rollen häufig verwendet werden, nicht aufgelöst (z. B. ›F. d.‹ oder ›T. d.‹, für ›Frau des‹ oder ›Tochter des‹). Die OhdAB nimmt sicher auflösbare Abkürzungen zudem als Schreibvarianten auf.

3.3.4 Definierte berufsferne Substantive

Direkte Angaben über den Rechtsstatus werden separiert, nicht aber aus der Berufsangabe entfernt. Hintergrund ist, dass diese Angaben Teil der Varianten der OhdAB sind und dadurch erkannt werden können. Das umfasst folgende Begriffe:

- Bürger
- Civis Academicus
- Einwohner
- Inwohner
- in wohner
- In wohner
- Nachbar
- Universitätsbürger

Angaben, die Auskunft über den Besitz geben, werden hingegen der Kategorie Besitzinformationen zugeordnet und aus der Berufsangabe gelöscht. Hierunter fallen alle von Leerzeichen umfassten Begriffe, die auf ›besitzer‹ oder ›besitzerin‹ oder ›eigentümer‹ und ›eigentümerin‹ enden.

3.3.5 Lokale Präpositionen

Ortsangaben können mit verschiedenen lokalen Präpositionen eingeleitet werden. Ist eine der folgenden Zeichenketten samt vorangehendem und nachfolgendem Leerzeichen Teil der Berufsangabe, so wird der nachfolgende Teil als Ortsangabe separiert und die Präposition gelöscht. Ein voranstehendes Leerzeichen ist nicht notwendig, wenn die Präposition am Beginn des Strings steht.

- in
- In
- i.
- von
- zu
- auf
- aus
- an
- der
- des

Abgegrenzt von der Ortsangabe weisen folgende Ergänzungen der Präposition ›bei‹ eher einen Bezug zu einem Arbeitgeber auf als zu einem physischen Ort. Hier wird die Kategorie *Arbeitgeber* verwendet.

- bei der
- bei dem

Des Weiteren werden folgende Adjektive, die keine lokale Präposition sind, ebenso in die Arbeitgeberkategorie separiert, aber nicht aus dem weiter zu verarbeiteten String gelöscht, da sie einen wichtigen Bestandteil für die weitere Klassifizierung darstellen und auch die Varianten der OhdAB diese Begrifflichkeiten mitführen.

- herrschaftlich
- herrschaftliche
- königlich

- königliche

3.3.6 Separierung von Quellenangaben

Verlinkungen werden in den Bereich der Quellen separiert und gelöscht. Hierunter fällt der Text zwischen <a> und (inklusive der beiden genannten Zeichen). Andere Quellenangaben werden nicht erkannt und erscheinen ggf. nachher in der Kategorie *Weiteres*.

3.3.7 Titelangaben

Falls die Berufsangabe Informationen zum Titel enthält, werden diese in die Kategorie *Titelangaben* separiert. Wenn auf eine der nachfolgenden Zeichenketten ein Leerzeichen folgt, so endet die Titelangabe mit dem Punkt. Eine Ausnahme besteht darin, dass der nachfolgend durch Leerzeichen abgetrennte Teilstring auch mit einem Punkt endet und somit eine Abkürzung darstellt. Hier wird auch dieser Teilstring in die Titelangabe mit eingebunden und gelöscht. Das betrifft auch weitere nachfolgende Teilstrings (z. B. ›Dr. rer. nat.‹). Folgt der Angabe ›Dr.‹ kein Leerzeichen, so sind alle Zeichen bis zum nächsten Leerzeichen zu separieren (z. B. ›Dr.iur.‹).

- Prof.
- Professor
- Dr.
- Herzog

Des Weiteren gibt es viele andere Titelangaben wie Titularherr, Graf, Contesse, Gräfin, Freifrau, Freiherr etc. Sie können nach Bedarf ergänzt werden.

3.3.8 Angaben zum Familienstand

Folgende Teilstrings werden in die Kategorie *Familienstand* separiert und gelöscht:

- F. d.
- Ehefrau des
- Ehefrau d.
- Ehefrau
- -frau (am Ende einer Bezeichnung)
- T. d.
- -tochter (am Ende einer Bezeichnung)
- S. d.
- -sohn (am Ende einer Bezeichnung)
- ›Witwe‹ oder ›Witwer‹
- ›Wittib‹ oder ›Wittiber‹
- ›Jungfrau‹ oder ›Jungfer‹
- ›Junggeselle‹ oder ›Junggesell‹

Dieses bezieht sich nicht auf definierte Ausnahmen, in denen dieses String Teil der Berufsangabe ist (z. B. ›Dienstfrau‹, ›Arbeitsfrau‹). Folgende Adjektive, die möglichen Familienstandsangaben (aber auch anderen Substantiven) vorangestellt sind, werden ohne Separierung gelöscht:

- ›ehrbare‹ oder ›ehrbarer‹
- ›tugendsame‹ oder ›tugendsamer‹
- ›wohlgeachtete‹ oder ›wohlgeachteter‹
- ›geachtete‹ oder ›geachteter‹

3.3.9 Temporale Präpositionen und Ziffern

Zunächst wird der String auf die folgenden temporalen Präpositionen durchsucht. Werden diese gefunden, wird das nachfolgende, durch vor- und nachstehende Leerzeichen abgegrenzte Wort als Zeitangabe separiert und samt Präposition aus dem String gelöscht.

- am
- im Jahr

Zeitangaben sind aber insbesondere auch durch zusammenhängende Ziffern ohne einleitende Präposition dargestellt. Der String wird zunächst auf die Ziffern 0 bis 9 durchsucht. Bei genau vier aufeinanderfolgenden Ziffern wird eine Jahreszahl angenommen. Diese wird separiert und gelöscht. Sollte vor der Jahreszahl jedoch ein Punkt auftauchen, so werden allen Zeichen davor bis zum nächsten Leerzeichen gelöscht. Ausschließlich die Jahreszahl wird separiert, da eine zeitlich genauere Verortung nicht notwendig erscheint.

3.3.10 Erkennung von Rollenangaben

Es werden sechs Rollen unterschieden:

- Gehilfe oder -gehilfe / Knecht oder -knecht / Magd oder -magd / Helfer oder -helfer / Bursche oder -bursche
- Lehrling oder -lehrling / Geselle oder -geselle
- Macher oder -macherin (Grundform des Berufes)
- Meister oder -meister
- Obermeister oder -obermeister / Oberältester oder -oberältester
- Besitzer oder -besitzer / Eigentümer oder -eigentümer

Dieser Zusatz wird nur festgestellt und in der Kategorie *Rolle* gespeichert, bleibt aber in der Berufsbezeichnung erhalten, wenn er von der eigentlichen Berufsangabe nicht getrennt ist. Steht er frei, so wird er ans Ende des darauffolgenden von Leerzeichen umschlossenen Teilstring gestellt.

3.3.11 Berufsstatus

Folgende Wörter dienen als Signalwörter, aus denen sich Rückschlüsse auf den aktuellen Berufsstatus ziehen lassen. Sie werden gelöscht und in die Kategorie *Berufsstatus* separiert.

- ›pensionierte‹ oder ›pensionierter‹
- ›a. D.‹ oder ›a.D.‹
- ›i. R.‹ oder ›i.R.‹
- ›gewesene‹ oder ›gewesener‹[40]

Auch hier existieren zahlreiche weiter Signalwörter (u. a. ›Alt-‹, ›weiland‹, ›emeritiert‹, ›vormaliger‹, ›vormals‹, ›verstorbener‹, ›verabschiedeter‹, ›verrenteter‹, ›früherer‹, ›ehemaliger‹, ›ausrangierter‹, ›abgedankter‹). Auch diese können bei Bedarf ergänzt werden.

3.3.12 Separation von Angaben in Klammern

Es wird davon ausgegangen, dass die wesentlichen Berufsangaben nicht in Klammern stehen. Diverse mögliche Inhalte für Klammern wurden bereits in den vorherigen Schritten entfernt. Die übriggebliebenen Daten können nicht genau zugeordnet werden und werden aus diesem Grund ohne die Klammern in die Kategorie *Weiteres* separiert und gelöscht. Die Klammern selbst werden gelöscht.

[40] Hier ist zu beachten, dass die Angabe auch auf den zuvor erfolgten Tod des Stelleninhabers hinweisen kann.

3.3.13 Löschung von Sonderzeichen

Verbleibende Sonderzeichen, mit Ausnahme von Punkten, die durchaus Teil einer Berufsangabe sein können, werden gelöscht. Als Sonderzeichen werden all jene Zeichen definiert, die keine Zahlen oder Buchstaben sind. Falls vor, nach oder vor und nach den Sonderzeichen ein Leerzeichen steht, so wird stattdessen ein Leerzeichen eingesetzt.

3.3.14 Umsetzen der Kleinschreibung

Verbleibende Großbuchstaben werden durch ihre entsprechende Kleinschreibung ersetzt. Dies dient dazu, Differenzen in der Groß- und Kleinschreibung zu ignorieren.

Der übergebliebene String wird nochmals von Leerzeichen am Anfang und Ende bereinigt. Er enthält abschließend nun die bereinigte Variante der Berufsangabe und wird ebenso einer Kategorie (Beruf) zugeordnet. Die Bereinigung dieses Strings ist damit abgeschlossen und er kann der Berufsangabenklassifizierung unterzogen werden. Demzufolge werden Tippfehler an dieser Stelle nicht erkannt, können aber durch die nachfolgende Ähnlichkeitsanalyse erfasst werden.

Die Ausführung der Verarbeitung (workflow execution) erfolgt nicht iterativ, sondern einmalig.[41] Um den Algorithmus auf die Angaben anzuwenden, ist eine Vorbereitung der Daten notwendig: Die Berufe müssen als Liste vorliegen, da das Ziel in einer automatisierten Klassifizierung besteht, in der definitionsgemäß kein post-processing and controlling durch eine manuelle Kontrolle notwendig ist. Aus den Spezifika des Datensatzes kann nun die Anpassung des Quellcodes geboten sein.

3.4 Klassifizierung der Berufsangaben

Nach der Bereinigung sind den Berufsangaben trotzdem noch keine Berufe der OhdAB-Konkordanz zugeordnet. Die notwendige Lemmatisierung 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 (Lemmatisierung) 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.[42] Folgend werden Möglichkeiten aufgezeigt, die im Abschnitt zur Validierung (Kapitel 5) untersucht werden. Wenn eine bereinigte Berufsangabe mit einer Variante exakt übereinstimmt, wird die Berufsangabe dieser Variante zugeordnet. 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.

3.4.1 Levenshtein-Distanz absolut

Die Levenshtein-Distanz wird jeweils für die Berufsangabe und die Varianten berechnet; zur Verbesserung der Laufzeit wird ein Vergleich nur bei einer Übereinstimmung des ersten Buchstabens vorgenommen. Aus einer hohen Ähnlichkeit dieser beiden Strings resultiert eine geringe Distanz. Zeichenketten mit einer Distanz von 1 werden als ähnlich klassifiziert und ausgewählt. Die absolute Levenshtein-Distanz wird auch als *Leva* bezeichnet.

[41] Bei der Entwicklung des Algorithmus hat ein iteratives Vorgehen jedoch sehr wohl Raum eingenommen. Durch die Begutachtung des Klassifikationsergebnisses wurden weitere Anomalien entdeckt, die in den Algorithmus mit eingebaut wurden.
[42] Kirby et al. 2015, S. 58.

3.4.2 Levensthein-Distanz relativ

Da in einer längeren Zeichenkette auch mehrere Fehler oder Variationen vorkommen können, wird die Levenshtein-Distanz mit der Länge der zu überprüfenden Berufsbezeichnung in Beziehung gesetzt (Formel 4). Dabei wird hier nicht differenziert, ob solche Fehler Resulat von Lese- oder Schreibprozessen, mangelhafter OCR-Erkennung oder tatsächliche Schreibvarianten sind. Unterschreitet die relative Distanz einen bestimmten Wert, findet eine Zuordnung statt. Der hierfür zu unterschreitende Grenzwert wird in der Validierung bestimmt.

Formel 4: Levr(bi,vj)=Lev(bi,vj)Länge bi

3.4.3 Erweiterung der Abkürzungserkennung

In der Validierung werden zwei verschiedene Möglichkeiten der Abkürzungserkennung verglichen: Zum einen ist das der Algorithmus, wie er zuvor vorgestellt worden ist (Auflösung definierter Abkürzungen). Zum anderen aber wird eine Erweiterung dahingehend getestet, ob bei ausbleibender Ähnlichkeit zu den Varianten eine Ähnlichkeit mit einer Abkürzung besteht. Dadurch wird z. B. für die Berufsbezeichnung ›Preußischer Leutnant‹ und der Variante ›Preuß. Leutnant‹ eine Übereinstimmung festgestellt, obwohl die ursprüngliche Levenshtein-Distanz vergleichsweise hoch ist.

3.4.4 Ergänzung einer lernenden Komponente während der Lemmatisierung

Die lemmatisierte Berufsangabe kann nun als Schreibvariante eines Berufs ebenso mit in die Varianten eingehen. Dadurch wird die Zahl der Varianten erhöht und die Wahrscheinlichkeit gesteigert, neue Berufsangaben zu erkennen. Der Vorteil gegenüber einer reinen Erhöhung von Grenzwerten ist an einem Beispiel gut erkennbar: Die Levenshtein-Distanz zwischen ›Müllerˆ und ›Muellerˆ ist möglicherweise zu groß, obwohl es denselben Beruf beschreibt. Wird nun über ›Müller‹ zuvor aber die Variante ›Muller” erkannt, wird im nächsten Schritt auch ›Mueller‹ erkannt. Bei einer erlaubten Levenshtein-Distanz von 2 wäre ›Mueller‹ zwar direkt erkannt worden, ›Maler‹ aber ebenso. Der Nachteil dieses lernenden Vorgehens besteht in der Reproduktion von Fehlern durch falsch-positive (FP) Ergebnisse.

3.4.5 Ergänzung einer lernenden Komponente im Anschluss in einer weiteren Iteration

Statt die neuen Varianten kontinuierlich hinzuzufügen ist es auch möglich, nach einer einmaligen Bearbeitung alle nicht-lemmatisierten Berufsangaben erneut zu untersuchen. Vorteil hierbei ist, dass die Berufsangaben zu Beginn (ohne gelernte Varianten) nochmals mit den später gelernten Varianten verglichen werden. Hierbei sind viele Iterationen vorstellbar.

4. Programmtechnische Umsetzung

Der im vorherigen Abschnitt beschrieben Algorithmus kann wie in Abbildung 2 zu sehen grafisch dargestellt werden.

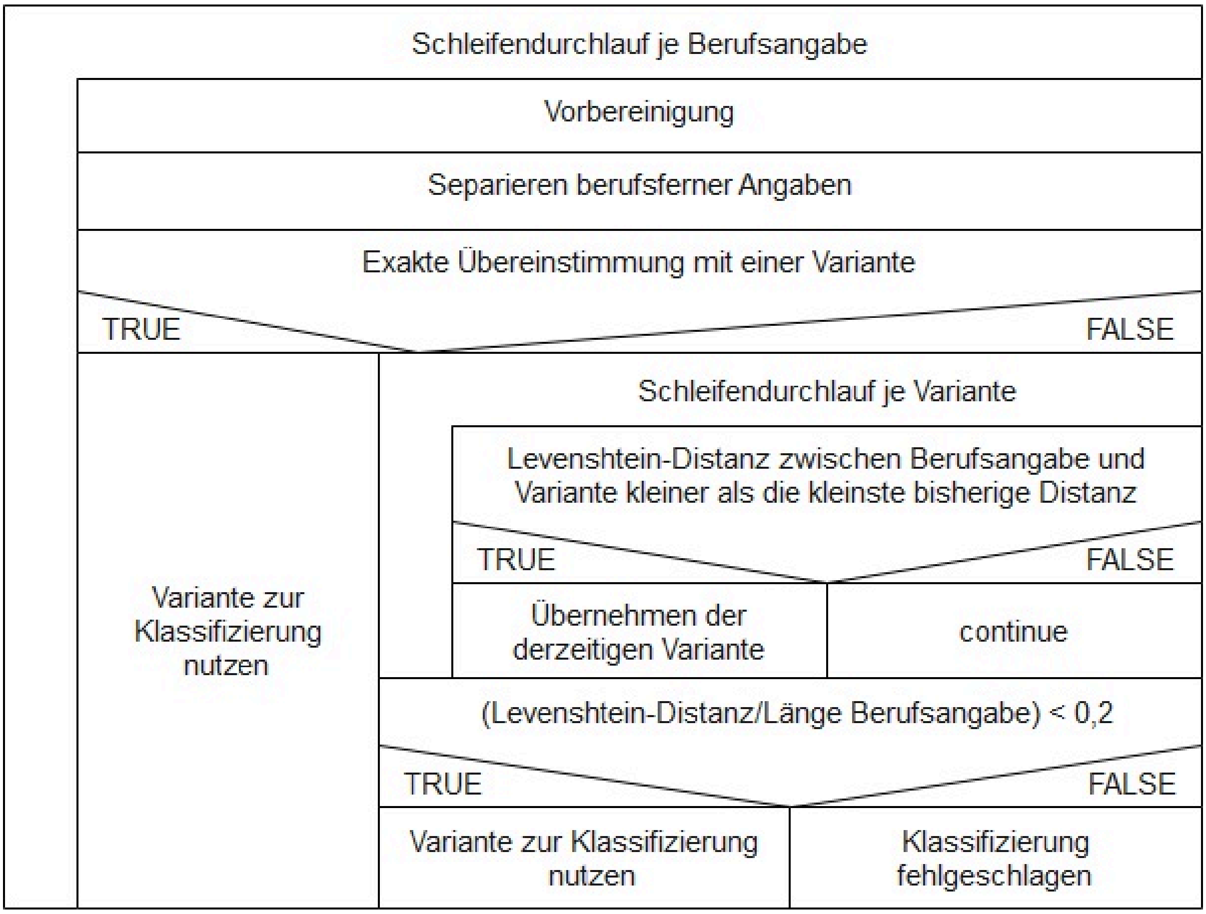

Abb. 2: Algorithmus, dargestellt in einem Nassi-Shneiderman-Diagramm. [Goldberg / Moeller 2022]

Zur Umsetzung des Algorithmus wird die Programmiersprache Python 3.7 verwendet. Diese bietet den Vorteil, dass für die Datenbereinigung keine dritte Software genutzt wird, die ggf. Lizenzeinschränkungen mit sich bringen würde. Das Ergebnis ist öffentlich zugänglich und kann für weitere wissenschaftliche Arbeiten verwendet oder angepasst werden. Dazu ist der Quellcode im Online-Repositorium zu finden. Er enthält die Variante des Algorithmus, die das beste Ergebnis in Bezug auf das F1-Maß erbringt (siehe folgender Abschnitt 5).

Das Programm ist in einzelne Funktionen gegliedert, welche im Folgenden vorgestellt werden, bevor das Zusammenwirken dieser erläutert wird. Die Vorstellung an dieser Stelle dient dazu, einen einfacheren Zugang zur Anpassung des Codes zu ermöglichen. Auf eine detaillierte Beschreibung der Funktionsweise wird an dieser Stelle verzichtet. Nähere Erläuterungen sind den Kommentaren im Programmcode zu entnehmen.

Der Abbildung 3 ist der grundlegende Aufbau des Programms zu entnehmen. Die Pfeile zwischen den Funktionen deuten darauf hin aus welcher übergeordneten Funktion diese aufgerufen werden. In der *main*-Funktion werden zunächst relevante Dateien ausgewählt, die dann parallelisiert über die Funktion `preCreateOccuList` aufgerufen werden (spätere Iterationen über `createOccuList`). Jede GEDCOM-Datei wird darin über die Funktion `loadGedcomFile` aufgerufen. Danach wird die Funktion `createOccuList` aufgerufen, in welcher ein Aufruf einer Liste bisheriger Varianten durch die Funktion `loadData` stattfindet. Über die Funktion `createFile` werden Ausgabedateien initial erstellt.

Mit der Funktion `occuCleaner` werden in der Funktion *createOccuList* die einzelnen Berufsangaben zunächst grundlegend bereinigt, die Schritte 1 bis 3 des Bereinigungsalgorithmus werden damit realisiert. Dazu werden Leerzeichen am Anfang und Ende entfernt und definierte Abkürzungen ausgeschrieben. Verschiedene Trennoperatoren werden zu ›und‹ normiert. Die Berufsangabe wird dann pro ›und‹ aufgespalten und in einzelne *Dictionaries* separiert. Die maximale Anzahl von Trennungen der Berufsangabe liegt hier bei 5. Dieser Separierung erfolgt in der Funktion `separator`.

Zu jeder vorbereinigten Berufsangabe wird nun die Funktion `partCorrector` ausgeführt, dort wird der Bereinigungsalgorithmus ab Schritt 4 umgesetzt. Ziel dieser Funktion ist es, pro Angabe ein Dictionary zu erzeugen, in dem die verschiedenen Bestandteile der Angabe dokumentiert werden. Das Dictionary enthält Informationen zur Berufsangabe, die Lemmatisierung dieser zu der OhdAB, mögliche vom Beruf abzugrenzende Titel, Rollen oder Ortsangaben sowie Zeitangaben und URLs. Alles, was in keine dieser Kategorien einsortiert werden kann, wird als *Weiteres* bezeichnet.

Um die Klassifizierung nach der OhdAB vornehmen zu können, wird in der Funktion `dictSearch` eine vollständige Übereinstimmung mit der bereinigten Berufsangabe geprüft. Besteht keine vollständige Übereinstimmung, so wird mithilfe der Levenshtein-Distanz (Funktion `levenshteinDist`) die Ähnlichkeit zu den anderen Varianten überprüft. Die Variante mit dem geringsten Wert bei dem Verhältnis von Levenshtein-Distanz und Länge der zu untersuchenden Berufsbezeichnung, wird ausgewählt. Bei gleicher Distanz wird die Variante ausgewählt, die von vorne beginnend die meisten übereinstimmenden Buchstaben mit der zu lemmatisierenden Bezeichnung aufweist. Liegt der Wert der relativen Levenshtein-Distanz unter 0,25 wird eine Übereinstimmung angenommen. Die Liste der Varianten selbst wurde über die Funktion `loadData` als Liste von Dictionaries hochgeladen. Dieses erklärt den Namen der Funktion `dictSearch`.

Die Ähnlichkeitsanalyse findet überwiegend in der Funktion `levenshteinDist` statt. Zur Auflösung von Abkürzungen wird zusätzlich die Funktion `abbreviationCorrector` verwendet. Um die Position bestimmter Teile in einem String zu ermitteln, wird die Funktion `endOfString` verwendet. Die Funktion `replaceLoc` hingehen dient der Separierung von Ortsbestandteilen aus der Bezeichnung. Der Zusammenhang der Funktionen ist in Abbildung 3 dargestellt.

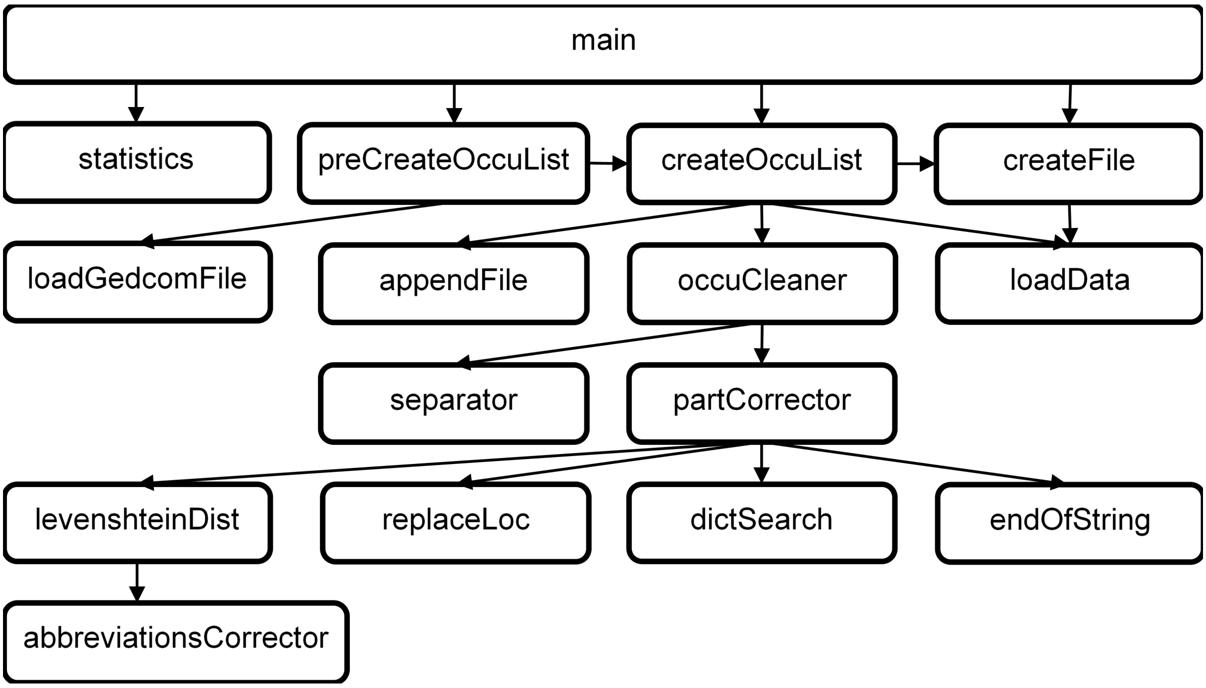

Abb. 3: Zusammenhang der Funktionen. [Goldberg / Moeller 2022]

5. Validierung und Diskussion

Zur Validierung werden zunächst 3,3 Prozent der Dateien ausgewählt (Trainingsdaten). In den zufällig ausgewählten 95 Dateien finden sich insgesamt 1.840 Berufsangaben. Diese werden zur Festlegung der Levenshtein-Distanz-Grenzen verwendet.
Die Reduzierung des Datenvolumens in diesem Schritt ist notwendig, da eine manuelle Überprüfung der Korrektheit der Klassifizierung aller Ergebnisse nur mit übermäßig viel Aufwand möglich wäre. Dabei wird zunächst geprüft, ob die absolute oder relative Levensthein-Distanz für den Algorithmus besser geeignet ist, und wie hoch der Grenzwert für eine Ähnlichkeitserkennung sein sollte. Anschließend daran wird geprüft, ob und wie die Abkürzungserkennung Einfluss auf das Ergebnis nimmt. Gleiches wird mit der erweiterten Bereinigung der Berufsangaben geschehen. Abschließend wird der Einfluss einer lernenden Komponente getestet, indem die neu erlernten Varianten in die Gesamtheit aller Varianten eingehen.

Da es Ziel des Algorithmus ist, das F1-Maß zu maximieren, ist festzulegen ab welchem Grenzwert – genannt *Leva* (Levenshtein-Distanz absolut) und *Levr* (Levensthein-Distanz relativ) – eine Zuordnung zwischen Berufsangabe und Variante erfolgen soll. Da die Bewertung, ob eine Zuordnung falsch oder richtig ist, hier nur manuell geschehen kann, ist eine Schätzung der Grenzwerte auf Basis aller Daten sehr zeitaufwändig und mit zunehmender Anzahl von Daten auch mit einem abnehmenden Grenznutzen für die Güte das Parameters behaftet. Zudem macht bei der relativen Levenshtein-Distanz ein grob gerundeter Parameter in den meisten Fällen keinen Unterschied. Beispielsweise bei einer achtstelligen Berufsangabe steigt *Levr* bei jeder Erhöhung der jeweiligen Levenshtein-Distanz um 0,125 (ein Achtel). Ob der Grenzwert im Beispiel also bei 0,126 oder 0,249 liegt ist irrelevant.

Zunächst zeigt die Durchführung einer Klassifikation mit der absoluten Levenshtein-Distanz mit einem Grenzwert von ≤ 1, ≤ 2 und ≤ 3 im Vergleich (siehe Tabelle 3), dass eine Distanz von 2 beziehungsweise 3 ein schlechteres Ergebnis in Bezug auf die Genauigkeit (P) erbringt. Dabei wird nur der Anteil der Berufsangaben in den Trainingsdaten herangezogen, die nicht durch einen genauen Treffer identifiziert werden, sodass nur die Berufsangaben übrigbleiben, bei denen die Ähnlichkeitserkennung einen Unterschied macht. Der Umfang dieser Berufsangaben an den Trainingsdaten ist jeweils den Spalten ›Anzahl‹ und ›Anteil‹ zu entnehmen.[43] Besonders deutlich wird die Ungenauigkeit bei einer absoluten Levenshtein-Distanz von 3, bei der lediglich etwa die Hälfte der Lemmatisierungen noch korrekt ist. Wenn jedoch angenommen wird, dass die Grundgesamtheit nur aus den 72 bei ≤ 3 erkannten Berufsangaben besteht, so kann ein F1-Wert berechnet werden. Hierbei ergibt sich ein maximaler Wert bei einer Levenshtein-Distanz von 2.

Lev	Anzahl	Anteil	TP	FP	P	FN	TN	R	F1
≤ 1	35	1,88 %	31	4	0,886	7	30	0,816	0,849
≤ 2	47	2,53 %	37	9	0,787	1	25	0,974	0,881
≤ 3	72	3,87 %	38	34	0,527	0	0	1,000	0,691

Tab. 3: Klassifikation unser Variation der Levenshtein-Distanz als Grenzwert. [Goldberg / Moeller 2022]

Bei einem Vergleich von verschiedenen Grenzwerten der relativen Levenshtein-Distanz zeigt sich zudem, dass ein Wert zwischen 0,25 und 0,30 die besten Ergebnisse erbringt (siehe Tabelle 4). Ein maximaler F1-Wert wird bei einer Grenze von *Levr* < 0,30 erzielt. Es zeigt sich zudem, dass die Genauigkeit (P) mit zunehmendem Grenzwert sinkt. Der Ausreißer der Genauigkeit bei < 0,3 ist eher dadurch bedingt, dass durch den Schritt von < 0,28 auf < 0,30 zufällig zwei weitere Berufsangaben der Stichprobe positiv und korrekt lemmatisiert werden. Wird angenommen, dass die Grundgesamtheit nur aus den 57 bei < 0,4 erkannten Werten besteht, so kann ein F1-Wert berechnet werden. Gute Werte ergeben sich zwischen < 0,2 und < 0,3. Das Maximum des F1-Werts wird durch die beiden Ausreißer zwar bei < 0,3 erreicht. In Hinblick auf das gute Ergebnis, das aber bereits bei < 0,2 erreicht wird, wird für die folgende Verarbeitung ein Grenzwert von < 0,25 ausgewählt.

Dieses Vorgehen ist allerdings nur bei Bezeichnungen mit einer Mindestlänge sinnvoll. Bei Wörtern unter fünf Buchstaben führt mehr als eine Änderung bereits zu einem Wert von 0,25 und somit niemals zu einer Zuordnung.[44] Eine weitere (erwünschte) Eigenschaft ist, dass bei vielen fremdsprachlichen Angaben, die als TN klassifiziert werden sollten, keine Zuordnung geschieht, da die relative Levenshtein-Distanz dort oftmals sehr hoch ist. Ebenso sieht es bei einer falschen Verwendung des Felder aus (z. B. Eintragung einer Ortsangabe). Vorteilhaft ist dieses Vorgehen insbesondere bei geringfügig differierenden Schreibvarianten, ausgelassenen oder zu viel vorhandenen Buchstaben sowie Buchstabendrehern.

Levr	Anzahl	Anteil	TP	FP	P	FN	TN	R	F1
< 0,10	10	0,54 %	10	0	1,000	27	20	0,270	0,426
< 0,20	34	1,85 %	31	3	0,912	6	17	0,837	0,873
< 0.25	37	2,01 %	33	4	0,892	4	16	0,891	0,892
< 0,28	39	2,12 %	33	6	0,846	4	14	0,891	0,868
< 0,30	41	2,23 %	35	6	0,878	2	14	0,946	0,897
< 0,40	57	3,10 %	37	20	0,649	0	0	1,000	0,787

Tab. 4: Klassifikation unser Variation des Grenzwerts einer relativen Levenshtein-Distanz. [Goldberg / Moeller 2022]

[43] Da ein großer Teil des F1-Maß durch die direkte Erkennung bestimmt ist und die Ähnlichkeitsanalyse nur einen kleinen Anteil ausmacht, wird hier nur der Teil der Daten betrachtet, der den Unterschied determiniert.
[44] Allerdings existieren nur wenige Berufsbezeichnungen unter fünf Buchstaben.

Nachteilig ist, dass Abkürzungen so nicht erkannt werden, da die absolute Levenshtein-Distanz zwischen einem Begriff und seiner Abkürzung definitionsgemäß mindestens die Anzahl der nicht vorhandenen, abgekürzten Buchstaben beträgt. Wird die Erkennung von Abkürzungen aktiviert, hat das auf die Trainingsdaten jedoch keine Auswirkung (getestet bei relativer Levenshtein-Distanz von < 0,2). Wird stattdessen eine Stichprobe von jeder zehnten Datei genommen (statt zuvor jeder fünften Datei), wird im Versuch eine weitere Berufsangabe gefunden (hier wird die Berufsangabe ›Landwirtschaftliche Arbeiterin‹ der Variante ›Landwirtschaftl. Arbeiterin‹ zugeordnet). Auch wenn solche Fälle (in den verwendeten Daten) nicht häufig vorkommen, so bleibt die Abkürzungserweiterung dennoch im Algorithmus, weil sie grundsätzlich die Güte des Ergebnisses verbessert.

Um den Einfluss der erweiterten Bereinigung der Berufsangaben auf die Güte des Ergebnisses zu prüfen, werden die Testdaten verwendet (229.669 Berufsangaben in 2.899 Dateien).[45] Hier werden nicht die Trainingsdaten verwendet, sondern alle Daten, weil vorrangig von Interesse ist, ob dadurch mehr Varianten gefunden werden. Bei einer Einbindung der Bereinigung können 64 Prozent der beruflichen Bezeichnungen direkt (ergo ohne Ähnlichkeitsanalyse) lemmatisiert und einer bestehenden Variante zugeordnet werden (siehe Tabelle 5). Das sind sieben Prozent mehr im Vergleich zu einem Durchlauf ohne diese Bereinigung. Bei den vergleichenden Bezeichnungen wird eine Ähnlichkeitsanalyse durchgeführt. Auch hier bringt die Bereinigung ein geringfügig besseres Ergebnis hervor (+0,22 Prozent Erkennung). Bei der Version mit Bereinigung bleiben 30 Prozent der Angaben über, die nicht erkannt werden können. Ein geringer Prozentsatz leerer Bezeichnungen ist auch enthalten, in denen keine Information zu finden ist. Wie hier auch zu sehen ist, hat die Ähnlichkeitsanalyse nur eine geringe Auswirkung im Vergleich zur direkten Erkennung. Diese wird durch den Einsatz der Bereinigung maßgeblich erhöht und stellt deshalb ein sehr wichtiges Element des Algorithmus dar.

	Direkt gefunden	Ähnlichkeitsanalyse	Nicht gefunden	Leere Bezeichnungen
mit Bereinigung (insgesamt 229.669 Angaben)				
Anzahl	147.781	9.674	68.955	3.259
Anteil	64,35 %	4,21 %	30,02 %	1,42 %
ohne Bereinigung (insgesamt 229.669 Angaben)				
Anzahl	131.064	9.160	86.344	3.101
Anteil	57,07 %	3,99 %	37,59 %	1,35 %

Tab. 5: Vergleich des Effektes der Bereinigung auf die Erkennung. [Goldberg / Moeller 2022]

Die durch die Ähnlichkeitsanalyse zugeordneten Berufsangaben können, da diese als Variante noch nicht existieren, in der Variantenliste ergänzt werden. Dieses kann auf zwei Arten geschehen: (1.) indem die neuen Treffer direkt nach Erkennung in die Menge der Varianten eingehen oder (2.) alle nicht erkannten Bezeichnungen im Anschluss nochmals mit allen neuen Varianten abgeglichen werden. Letzteres kann in mehreren Iterationen durchgeführt werden. Hierbei zeigt sich, dass die nachfolgende, zweifach-iterative Verarbeitung ein besseres Ergebnis in Bezug auf das F1-Maß ergibt als die kontinuierliche Ergänzung (siehe Tabelle 6).[46] Zwar kann bei dieser Option eine niedrigere Genauigkeit (P) beobachtet werden, doch sorgt die große Anzahl zusätzlich erkannter Angaben für eine Steigerung des F1-Wertes. Es ist anzunehmen, dass eine hohe FP-Rate bei den Iterationen der Ähnlichkeitserkennung tendenziell zu einer Fortführung von Fehlern führen kann, weswegen viele zusätzliche Iterationen nicht sinnvoll erscheinen.

Dabei ist zudem anzunehmen, dass der Lerneffekt größer ist, je mehr Berufsangaben verarbeitet werden, da die Chance steigen könnte, dass eine ähnliche Bezeichnung auftritt. Bei einem exemplarischen Durchlauf mit jeder zehnten Datei wird noch keine zusätzliche Erkennung erreicht. Auch bei einer Verarbeitung mit allen Daten werden nur weitere 0,01 Prozent der Berufsangaben dadurch zusätzlich lemmatisiert. Dieser geringe Wert ist darauf zurückzuführen, dass bereits sehr viele Schreibversionen in den zugrundeliegenden Varianten der OhdAB abgedeckt sind. Bei einer zufälligen Halbierung der in der OhdAB vorhandenen Varianten steigt der Anteil der so zusätzlich erkannten Angaben deutlich um 8,80 Prozent (von 4,21 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.

[45] Das entspricht den Dateien, die nicht in den Trainingsdaten vorhanden sind.
[46] Von den durch die Ähnlichkeitsanalyse erkannten Daten werden 100 zufällige Werte manuell überprüft. Durch diese wird auf die Rate an TP- und FP-Werte geschlossen. Um einen F1-Wert zu berechnen ist zusätzlich die Anzahl von FN-Werten notwendig. Wie bereits zuvor wird dabei von der maximalen Anzahl erkannter Angaben ausgegangen (hier bei der zweifachen Iteration).

Verfahren	Anzahl	Anteil	TP-Rate in %	FP-Rate in %	P	FN	R	F1
Analyse mit sämtlichen ursprünglichen Varianten								
Ohne Lernen	9.674	4,21 %	88	12	0.88	5.943	0,59	0,71
Kontinuierlich lernen (4x Multiprocessing[47])	10.128	4,41 %	86	14	0.86	5.489	0,61	0,71
Iterativ lernend (1x)	11.185	4,87 %	83	17	0.83	4.432	0,68	0,75
Iterativ lernend (2x)	15.617	6,80 %	83	17	0.83	0	1,00	0,91
Analyse unter zufälliger Halbierung der ursprünglichen Varianten								
Ohne Lernen	27.583	12,01 %	80	20	0,80	6.086	0,78	0,79
Kontinuierlich lernen (4x Multiprocessing)	27.882	12,14 %	86	14	0,86	5.787	0,81	0,83
Iterativ lernend (1x)	32.774	14,27 %	76	24	0,76	895	0,97	0,85
Iterativ lernend (2x)	33.669	14,66 %	83	17	0,83	0	1,00	0,91

Tab. 6: Vergleich der Ähnlichkeitsanalyse unter Variation des maschinellen Lernens und unter Halbierung der zugrundeliegenden Berufsvarianten der OhdAB. [Goldberg / Moeller 2022]

Durch den Algorithmus – und dessen programmtechnische Umsetzung – wird in der Folge eine automatisierte Lösung zur Lemmatisierung deutschsprachiger Berufsangaben geboten. Insgesamt wird das F1-Maß optimiert, wenn eine relative Levenshtein-Distanz gewählt wird, Abkürzungen erweitert werden, eine Bereinigung stattfindet und erlernte neue Varianten im Anschluss nochmal mit allen Daten verglichen werden, die nicht lemmatisiert werden konnten. Ohne die Halbierung der Varianten, unter Herausrechnung der leeren Berufsangaben und mit doppelter Iteration des maschinellen Lernens wird eine Erkennungsrate von 72,17 Prozent erzielt (65,27 Prozent direkt und 6,90 Prozent über die Ähnlichkeitsanalyse). Die Halbierung der Varianten erhöht zwar den Anteil der über die Ähnlichkeitsanalyse erkannten Angaben, verringert jedoch die Zahl der direkt gefundenen Treffer. Herausfordernd ist für den Algorithmus vor allem auch, dass die GEDBAS-Daten sehr schwierig zu klassifizieren sind, weil eben nicht nur einfache Berufe angegeben werden. Es ist anzunehmen, dass mit qualitativ hochwertigeren Berufsangaben die Erkennung noch besser funktionieren würde, sodass hier ein grober Wert von 72 Prozent Erkennungsrate für diesen Algorithmus angegeben wird. Die Angabe, dass 98 Prozent der erkannten Werte auch korrekt sind, basiert darauf, dass alle direkt erkannten Werte als richtig bewertet werden. Zudem zeigt Tabelle 6, dass bei der angewendeten Spezifizierung der Ähnlichkeitsanalyse mit einer FP-Rate von 17 Prozent zu rechnen ist. Daraus ergibt sich eine TP-Rate über alle erkannten Berufe von etwa 98 Prozent.[48] Zudem ist es durch den Algorithmus möglich, berufsferne Angaben von der eigentlichen Bezeichnung des Berufs zu separieren. Der Algorithmus ist offen zugänglich und wird damit der Community zur Weiternutzung zur Verfügung gestellt. Es ist wünschenswert, dass er auch in anderen Anwendungen implementiert und stetig verbessert wird.

[47] Hierfür wurde die Parallelisierung mit vier Prozessorkernen verschiedenen Strängen ausgeführt. Das hat die Auswirkung, dass die Erkennung in einem Strang auf einen parallel ausgeführten keine Auswirkung hat (bei einer nachfolgenden Ausführung sich ggf. aber ausgewirkt hätte).
[48] Berechnung der TP-Rate: (65,27*1+6,90*0,83) / (65,27+6,90) = 0.98.

6. Zusammenfassung

Variationen einer Berufsbezeichnung können in der vorgestellten Weise automatisiert einem normierten Beruf zugewiesen werden. Das ermöglicht insbesondere der wirtschafts- und sozialhistorischen Forschung eine schnelle Klassifizierung großer Datenbestände, die für eine Vielzahl weiterer Anwendungen bereitsteht. Der entwickelte Algorithmus stellt eine Methode dar, mit der eine automatisierte Klassifizierung von historischen Standes- und Berufsangaben in einer hohen Güte vorgenommen werden kann: Von etwa 230.000 getesteten Berufsangaben aus der genealogischen Datenbank GEDBAS konnten rund 72 Prozent einem Beruf zugeordnet werden, wovon der wesentliche Teil von 98 Prozent auch korrekt ist. Dieses wird ermöglicht durch:

1. die Implementierung einer Bereinigung der Berufsangabe
2. eine Ähnlichkeitsanalyse zu bereits klassifizierten Schreibvarianten
3. die Implementierung einer Erweiterung von Abkürzungen und
4. eine Möglichkeit des überwachten maschinellen Lernens auf Basis der Treffer aus der Ähnlichkeitsanalyse

Jedes dieser Elemente führt zu einer Verbesserung des Ergebnisses. Das ist vor dem Hintergrund vieler fremdsprachlicher Bezeichnungen sowie einer sehr individuellen Eintragung der Berufsangaben in den GEDBAS-Daten ein zufriedenstellendes Ergebnis.

Dadurch, dass die Lemmatisierung auf den Daten der OhdAB aufbaut, der das Klassifizierungssystem KldB 2010 zugrunde liegt, ist sie besonders für das deutschsprachige Umfeld von Berufsbezeichnungen seit dem 16. / 17. Jahrhundert geeignet. Nach der Standardisierung mit der OhdAB sind zudem transparent abbildbare Neuansetzungen zeitspezifischer Klassifikationen möglich. Der Algorithmus kann jedoch auch als Ausgangspunkt genutzt werden, um ihn auf andere Sprachen anzupassen. Für die Begriffe der KldB 2010 existiert beispielsweise eine englische Übersetzung. Wenigstens auf der Ebene der Klassifikation würden vermutlich gute Ergebnisse produziert werden können. Herausforderungen liegen hierbei eher in der Schaffung der grundlegenden Datenbasis für die Indivdualbezeichnungen (Varianten) der Berufe. Neben der Einbindung der nicht-deutschsprachigen Varianten ist auch hier eine Anpassung der Anomalien im Algorithmus von großer Relevanz. Möglicherweise ergibt eine Abgrenzung einzelner Sprachen Sinn, damit keine ungewollten Übereinstimmungen in einem sprachenübergreifenden Programm auftreten.

Aber auch bei der Anwendung an deutschsprachigen Berufsangaben kann eine Anpassung des Programms helfen: Besondere Anomalien in den zu klassifizierenden Daten (z. B. spezifische Abkürzungen) können die Qualität des Ergebnisses für eine spezifische Anwendung verbessern. Des Weiteren können zusätzliche Verfahren der Berufsklassifizierung integriert werden (z. B. HISCO). Für den Algorithmus ist es allerdings von Vorteil, möglichst viele Variationen der Schreibweisen eines Berufes in dem jeweiligen System bereits klassifiziert zu haben. Zudem ist es denkbar, den Algorithmus nicht nur auf zuvor separierte Berufsangaben anzuwenden, sondern dahingehend zu erweitern, Berufsangaben in Fließtexten zu erkennen und auszulesen. Denkbar ist eine Einbindung von OhdAB in Verfahren der *Named Entity Recognition*, die auf Vokabularen aufsetzen.

Bibliografische Angaben

Arbeitskreis für Wirtschafts- und Sozialgeschichte Schleswig-Holsteins: Berufe in Altona 1803. Berufssystematik für eine präindustrielle Stadtgesellschaft anhand der Volkszählung. Kiel 1991. (= Kleine Schriften des Arbeitskreises für Wirtschafts- und Sozialgeschichte Schleswig-Holsteins, 1). [Nachweis im GVK]

Adam Friedrich Böhme: Anleitung wie Kirchenbücher zweckmäsig und ordentlich einzurichten sind. Leipzig 1790. [online] [Nachweis im GVK]

Marcel Bollmann: A Large-Scale Comparison of Historical Text Normalization Systems. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Bd. 1:_ Long and Short Papers. Hg. von Association for Computational Linguistics. (NAACL 2019, Minneapolis, MN, 02.07.–07.07.2019). Minneapolis, MN, Juni 2019, S. 3885–3898. PDF. DOI: 10.18653/v1/N19-1389

Klassifikationen der Berufe - Statistik der Bundesagentur für Arbeit. Hg. von Bundesagentur für Arbeit. Nürnberg 2021. [online]

Klassifikation der Berufe, Hg. von Bundesagentur für Arbeit. Nürnberg 2010. Bd 1 (2011): Systematischer und alphabetischer Teil mit Erläuterungen. [Nachweis im GVK]

Peter Christen / Karl Goiser: Quality and Complexity Measures for Data Linkage and Deduplication. In: Quality Measures in Data Mining. Hg. von Fabrice Guilett / Howard J. Hamilton. Berlin 2007, S. 127–151. [Nachweis im GVK]

Church of Jesus Christ of Latter-day Saints: The GEDCOM Standard. Release 5.5.1. 2019. PDF. [online]

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

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

Miguel Domingo / Francisco Casacuberta: Two Demonstrations of the Machine Translation Applications to Historical Documents. 02.02.2021. PDF. DOI: 10.48550/arXiv.2102.01417

Michael Dunn: Language phylogenies. In: The Routledge Handbook of Historical Linguistics. Hg. von Claire Louise Bowern / Bethwyn Evans. London u. a. 2015, S. 190–192. [Nachweis im GVK]

Tom Fawcett: An introduction to ROC analysis. In: Pattern Recognition Letters. In: ROC Analysis in Pattern Recognition 27 (2006), H. 8, S. 861–874. [Nachweis im GVK]

Corry Gellatly: Reconstructing Historical Populations from Genealogical Data Files. In: Population Reconstruction. Hg. von Gerrit Bloothooft et al. Cham 2015, S. 111–128. [Nachweis im GVK]

Metzler Lexikon Sprache. Hg. von Helmut Glück. 2., überarbeitete und erweiterte Auflage. Stuttgart u. a. 2000. [Nachweis im GVK]

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

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

Andreas W. Hauser / Klaus U. Schulz: Unsupervised Learning of Edit Distance Weights for Retrieving Historical Spelling Variations. In: Finite-state Techniques and Approximate Search. International Workshop. Hg. von Stoyan Mihov / Klaus U. Schulz. (International Workshop, International Conference RANLP 2007, Borovets, BG, 27.09.–29.09.2007). Borovets, BG, 30.09.2007, S. 1–6. PDF. [online]

Paul Hinschius: Das preußische Gesetz über die Beurkundung des Personenstandes und die Form der Eheschließung vom 9. März 1874 mit Kommentar in Anmerkungen. Berlin 1874. [Nachweis im GVK]

Kasra Hosseini / Federico Nanni / Mariona Coll Ardanuy: DeezyMatch: A Flexible Deep Learning Approach to Fuzzy String Matching. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Hg. von Association for Computational Linguistics. (EMNLP 2020, online, 16.11.–20.11.2020). Oktober 2020, S. 62–69. PDF. DOI: 10.18653/v1/2020.emnlp-demos.9

ISCO - International Standard Classification of Occupations. Hg. von ILO. Genf 2021. [online]

1306 records in total. Hg. von International Institute of Social History. In: History of Work Information System. Leuven 2020. [online]

Bryan Jurish: Finite-state Canonicalization Techniques for Historical German. Dissertation, Universität Potsdam. Potsdam 2012. PDF. [online]

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. In: Population Reconstruction. Hg. von Gerrit Bloothooft / Peter Christen / Kees Mandemakers / Marijit Schraagen. 2015, S. 43–60. DOI: 10.1007/978-3-319-19884-2

Jürgen Kocka / Claus Offe / Beate Redslob: Geschichte und Zukunft der Arbeit. (Konferenz, Berlin, 04.–06.03.1999) Frankfurt/Main 2000. [Nachweis im GVK]

Martin Kohli: Die Institutionalisierung des Lebenslaufs. Historische Befunde und theoretische Argumente. In: Kölner Zeitschrift für Soziologie und Sozialpsychologie 37 (1985), H. 1, S. 1–29. [Nachweis im GVK]

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

Marco H. D. van Leeuwen / Ineke Maas / Andrew Miles: History Of Work Information System. In: HISCO. Historical International Standard Classification of Occupations. Hg. von IISH / Antenna. Leuven 2002. [online]

Vladimir Iosifovič Levenštejn: Binary Codes Capable of Correcting Deletions, Insertations, and Reversals. In: Soviet Physics - Doklady 10 (1966), S. 707–710. [Nachweis im GVK]

Katrin Moeller: 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 / Anne Purschwitz. Halle 2019, S. 17–43. [Nachweis im GVK]

Katrin Moeller / Andreas Müller / Robert Nasarek: Ontologie historischer, deutschsprachiger Berufs- und Amtsbezeichnungen. In: geschichte.uni-halle.de/struktur/hist-data/ontologie/. Hg. von Historischen Datenzentrums Sachsen-Anhalt. Halle 2020. Beitrag vom 25.11.2020. [online]

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

Wiebke Paulus / Britta Matthes: 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 2013. PDF. [online]

Michael Piotrowski: Natural Language Processing for Historical Texts. San Rafael, 2012. (= Synthesis Lectures on Human Language Technologies, 17). [Nachweis im GVK]

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

Udo Schäfer: Die Novellierung des Personenstandsgesetzes. In: Archive, Familienforschung und Geschichtswissenschaft: Annäherungen und Aufgaben. Hg. von Bettina Joergens / Christian Reinicke. Düsseldorf 2006, S. 122–136. [Nachweis im GVK]

Reinhard Schüren: Soziale Mobilität. Muster, Veränderungen und Bedingungen im 19. und 20. Jahrhundert. St. Katharinen 1989. [Nachweis im GVK]

Derek Tam / Nicholas Monath / Ari Kobren / Aaron Traylor / Rajarshi Das / Andrew McCallum: Optimal Transport-based Alignment of Learned Character Representations for String Similarity. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Hg. von Association for Computational Linguistics. (ACL 57, Florenz, 28.07.–02.08.2019). Florenz, Juli 2019, S. 5907–5917. PDF. DOI: 10.18653/v1/P19-1592

Thesaurus Professionum. In: online.uni-marburg.de/fpmr/thepro/rs.php. Hg. von Universität Marburg: Forschungsstelle für Personalschriften an der Philipps-Universität Marburg. Marburg 2021. [online]

GEDBAS: Statistics. In: gedbas.genealogy.net/statistic/index. Hg. von Verein für Computergenealogie e. V. Köln 2021. [online]

Abbildungs- und Tabellenverzeichnis