Namenlemmatisierung in der Web-Datenbank mittelalterlicher und frühneuzeitlicher Universitätsmatrikeln

Stefan Böttcher, Rita Hartel, Heike Hawicks

DOI: 10.2436/15.8040.01.33

Abstract

Im Rahmen der Erstellung des Prototyps einer Web-Datenbank mittelalterlicher und frühneuzeitlicher Universitätsmatrikeln durch die Universitäten Duisburg-Essen und Paderborn sind bei der Namenrecherche insbesondere variierende Schreibweisen zu beachten, da die Schreibungen unstandardisiert aus dem Original übernommen werden, um die Einträge auch für namenkundliche Fragestellungen zu öffnen. Ausgehend von den Erfahrungen bei Datenbankprojekten mit frühmittelalterlichen Personennamen, deren extrem uneinheitliche Schreibvarianten durch zwei Lemmatisierungsansätze erschlossen wurden, sollen bei dem spätmittelalterlich-frühneuzeitlichen Namenmaterial automatisierte Namenausgleichsverfahren erprobt und eingesetzt werden. Neben der Lemmatisierung kommen auch phonetische Suchverfahren und Ansätze zum Data Cleaning zum Einsatz. Diese ermöglichen nicht nur, ähnliche oder verwandte Namen oder Begriffe zu finden, sondern bieten auch eine gewisse Fehlertoleranz sowohl bei der Erfassung, als auch insbesondere bei der Suche nach Namen. Bei diesen phonetischen Suchverfahren sollen unterschiedliche Algorithmen zunächst erprobt und dann sukzessive verbessert werden, um sie für ihre Anwendung bei der Datenbank der Universitätsmatrikeln optimal nutzbar zu machen. Das dabei auf phonetischer Grundlage (weiter)entwickelte Lemmatisierungsverfahren soll etymologisch gleiche Namen aufspüren, auch wenn sich die Namen durch einzelsprachliche Entwicklungen in der Orthographie und im Lautbild mitunter deutlich unterscheiden. Sein Anwendungsgebiet kann über die Erschließung historischer Quellen wie Matrikellisten hinausgehen, wenn sich diese Verfahren auch auf andere Namencorpora sowie die Erfassung von Daten aus handschriftlich ausgefüllten Formularen in herkömmlichen Datenbanken anwenden lassen, bei denen Fehler bei der Datenerfassung zu kompensieren sind. Schließlich eignen sich die Verfahren dazu, auch die Suche mit „unscharfen“ oder orthographisch fehlerhaften Suchbegriffen zu unterstützen, was bei historischen wie modernen Schreibungen von Vorteil ist, wenn man Überlieferungsverluste an Originalen oder die Problematik der Wiedererkennungsfähigkeit von Software beim Scannen in die Überlegungen einbezieht.

*****

Das Matrikel-Projekt, das an der Universität Duisburg begonnen wurde und nun mit den Matrikeln der Universität Heidelberg seine Fortsetzung findet, wurde inspiriert von dem Projekt Nomen et Gens, in dem es um frühmittelalterliche Namen und Personen geht.

In jenem Projekt, das von der Deutschen Forschungsgemeinschaft gefördert wurde, steht als Grundlage eine seit nunmehr über 30 Jahren in akribischer Quellenarbeit zusammengetragene Datenfülle von ca. 460.000 Datenbankeinträgen mit Personennamenbelegen zur Verfügung, welche sich aus urkundlichen Zeugenreihen, Gedenkbüchern und Necrologien speist. Das Material befindet sich in zwei Datenbanken – der seit Ende der 60er Jahre entwickelten DMP (Datenbank mittelalterlicher Personennamen) mit ca. 400.000 (zuletzt Runde/Voigt, 2002, 1252f.) und der Ende der 90er Jahre neu initiierten NeG (Nomen et Gens)-Datenbank mit ca. 60.000 Einträgen (Geuenich/Runde, 2006, passim). Der Hauptunterschied zwischen beiden besteht darin, dass die NeG-Datenbank nun relational aufgebaut ist, wohingegen die DMP noch aus Datensätzen in je einem String besteht.

Um die stark differierenden Schreibungen von frühmittelalterlichen Namen bzw. Namengliedern so zu ordnen, dass Personen aus ihnen herausgearbeitet werden können, wurde schon vor einigen Jahrzehnten ein Lemmatisierungsverfahren entwickelt, welches zuletzt in Uppsala 2002 Thema eines ICOS-Vortrags gewesen ist (Geuenich/Hawicks, 2008). Ziel eines übergeordneten Lemmas ist es, darunter möglichst viele Schreibungen eines Namens zu versammeln, mit denen möglicherweise ein und dieselbe Person bezeichnet

worden sein kann. Für das Frühmittelalter waren die Namenglieder der meist zweigliedrigen Personennamen Ausgangspunkt der Lemmata.

Als Beispiele seien zwei Namen aufgeführt, an denen Relevanz und Problematik der Lemmatisierung bzw. des entsprechenden Programms aufgezeigt werden können: Lothar und Eckart. Bei Letzterem greift schon ein erster Blick ins Internet-Lexikon Wikipedia die hier diskutierte Problematik auf. Dort heißt es: „Bei der Schreibweise seines Namens weisen die Handschriften die unterschiedlichsten Varianten auf (wie Aycardus, Ekhartus oder Hechard).“ (http://de.wikipedia.org/wiki/Benutzer:Eckhart_Triebel/Meister_Eckhart, 28.11.2011)

Hier folgen einige Schreibvarianten für das Erstglied des Namens, das Lemma „agi“, aus dem Lexikon der DMP-Datenbank: *acca, aekki, ecca, echi, egge, ehcc, exte, haeck, hecchi, heig* etc. pp.

Grundlage ist hier möglicherweise ein auf die Wurzel **agja* zurückzuführendes ahd. *ekka* im Sinne von 'Schwertesschärfe' (Förstemann, 1900, 14ff.; Kaufmann, 1968, 20ff.), welches mit "hardu" aus got. *hardus*, ahd. *hart* mit der Bedeutung 'kräftig', 'tüchtig', 'kühn' (Förstemann, 1900, 749ff.; Kaufmann, 1968, 173f.) kombiniert wurde.

Ebenso stellt sich die Variationsbreite bei Lothar dar, der ebenso als Chlothar/Chluthar, Hlodhar/Hludher, Lutar -> Luther oder gar Flothar/Floterus erscheinen kann.

Entsprechend folgen hier einige Schreibvarianten für das Erstglied des Namens, das Lemma „hlud“ aus dem Lexikon der DMP-Datenbank: *chlod, flodo, glood, hloud, hluda, hluudu, lauda, liudo, luod, lutthe* etc. pp.

Zugrunde liegt hier "hloda" aus der Wurzel **hlu* (hören), welches im Deutschen dem ahd. *hlut* im nhd. Sinne von laut am nächsten steht (Förstemann, 1900, 848ff.; Kaufmann, 1968, 189ff.). Das Zweitglied speist sich aus "harja", ahd. *har* = Heer bzw. Volk (Förstemann, 1900, 760ff.; Kaufmann, 1968, 174f.).

Doch wie bringt man diese zahllosen in frühmittelalterlichen Handschriften kursierenden Varianten unter einem verbindlichen Stichwort zusammen, denn ein Computer erkennt diese Gemeinsamkeiten, die ein Mensch mit entsprechenden Vorkenntnissen sofort durchschaut, wohl kaum. Hier hilft nur ein vordefiniertes Lemma, das nach philologischen Prinzipien geschaffen und als verbindlich festgelegt wird, und unter dem die Schreibvarianten allesamt versammelt werden. Dies wurde im DFG-Projekt Nomen et gens mit einer halbautomatischen und anschließenden handverlesenen philologischen Lemmatisierung praktiziert (Geuenich/Hawicks, 2008).

Mit dem Wechsel von der Einnamigkeit zur Zweinamigkeit im hohen Mittelalter, welcher schon häufig Gegenstand namenkundlicher Betrachtung war (Geuenich, 2002, mit Literatur), verändert sich die Anforderung an die Lemmatisierung von Personennamen grundsätzlich. Der Namenschatz wurde insgesamt geringer (Mitterauer, 1993, 241ff.) und die verhältnismäßig kleinere Zahl an Vornamen führte zu der Notwendigkeit, Zweitnamen zu verwenden, welche sich im Laufe der Zeit zu unveränderlichen Personen- bzw. Familiennamen verfestigten (Koß, 2002, 38ff.). Einzig die Schreibung dieser Vor- und Zunamen variierte noch, jedoch lässt sich diese Variation in keinster Weise mit dem Variantenreichtum der Namenglieder des Frühmittelalters vergleichen. Die Anforderung an eine Lemmafindung bzw. die Zuordnung der Schreibung an ein Lemma für einen hoch- und spätmittelalterlichen bzw. frühneuzeitlichen Namen ist somit eine gänzlich andere als im Frühmittelalter.

An den Matrikeln mittelalterlicher Universitäten lassen sich diese Probleme bestens studieren, wobei sich dieses umfangreiche wie variantenreiche Korpus geradezu dazu

anbietet, Lösungswege für die Personensuche und -identifikation zu erproben und bestenfalls zu finden. Diese Aufgabe hat sich das vorgestellte Projekt gestellt. Dazu bedarf es der Zusammenarbeit von an den Quellen arbeitenden Historikern resp. Archivaren, Sprachwissenschaftlern, Namenkundlern und EDV-Spezialisten.

Da die Thematik der Personensuche und -identifkation auch für heutige Fragestellungen relevant ist, wenn man z.B. an spezielle Recherchen in Einwohnermelderegistern oder bspw. Telefonbüchern etc. denkt, haben sich auch schon einige Forscher damit beschäftigt, Suchmotoren zu entwickeln, die wegen der grundsätzlichen lautlichen Nähe der Schreibvarianten eines Namens phonetische Algorithmen zugrundelegen. Im Rahmen dieses Beitrags sollen die Beispiele Soundex (1918) und Metaphone (1990) sowie die Kölner Phonetik (1969) vorgestellt werden.

Dabei legen alle Algorithmen den Schwerpunkt auf den Konsonantismus und nicht auf den Vokalismus, da letzterer in Namen die größte Stabilität aufweist. Die größte Varianz an Schreibungen findet sich hingegen bei den Konsonanten eines Namens, so dass diese in ähnlich lautende Gruppen zusammengefasst werden. Um diese Gruppen zu verstehen, sollen hier vorab einige Grundlagen der Phonetik, und zwar Artikulationsart und -ort erklärend zur Erläuterung der Gruppenbildungen bildhaft vor Augen geführt werden werden.

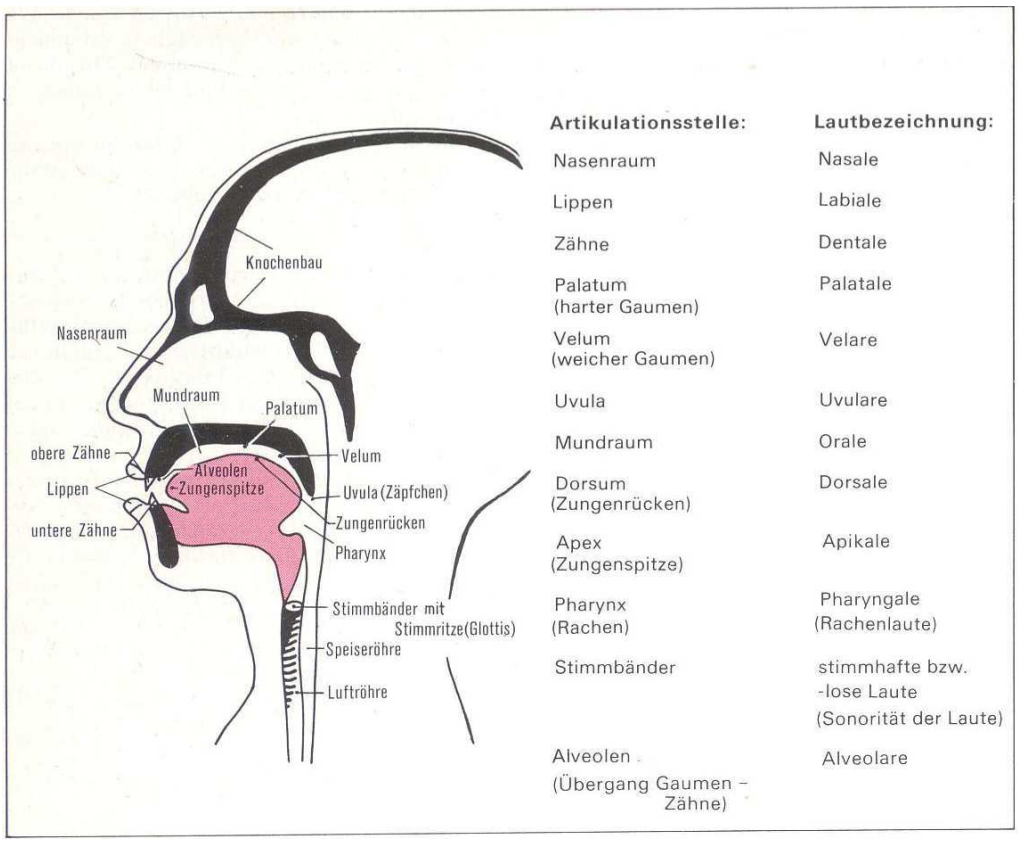


Abbildung 1: Schema des menschlichen Sprechapparates (König, 1985, 16).

Soundex (Russell, 1918) wurde 1918 von Robert Russell und Margaret Odell in den USA entwickelt. Der Soundex-Code eines Wortes besteht aus dem ersten Buchstaben des Wortes, gefolgt von Ziffern, die die Laute der nach dem Anfangsbuchstaben folgenden Konsonanten des Wortes repräsentieren. Hierbei werden ähnlich klingende Buchstaben durch den gleichen Code repräsentiert. Tabelle 1 zeigt die Zuordnung der Buchstaben einer Lautgruppe zum jeweiligen Soundex-Code.

Buchstabe	Code-Ziffer
A, E, H, I, O, U, W, Y	0
B, F, P, V	1
C, G, K, Q, S, Z	2
D, T	3
L	4
M	5
N	6
R	7

Tabelle 1: Zuordnung von Lautgruppen zu Soundex-Codes.

Bei Soundex werden (wie bei allen Algorithmen dieses Typs) die Vokale weitgehend ignoriert und können daher in einer Gruppe zusammengefasst werden (s. Code-Ziffer 0).

W und H werden bei Soundex zu dieser Gruppe hinzugezogen (bei der Kölner Phonetik wird dies anders gehandhabt). Die Gemeinsamkeit der Konsonanten, die unter Code 1 zusammengefasst werden (BFPV) liegen im Artikulationsort, da es sich um Bilabiale (b,p) bzw. Labiodentale (f,v) handelt (s. Abbildung 2). Code 2 fasst velare Verschlusslaute (g,k) und dental/alveolare Reibelaute (s,z) zusammen Die dental/alveolaren Verschlusslaute hingegen bilden eine eigene Gruppe (d,t). L und r als liquide Buchstaben bilden eine je eigene Gruppe, ebenso die beiden Nasale (m und n).

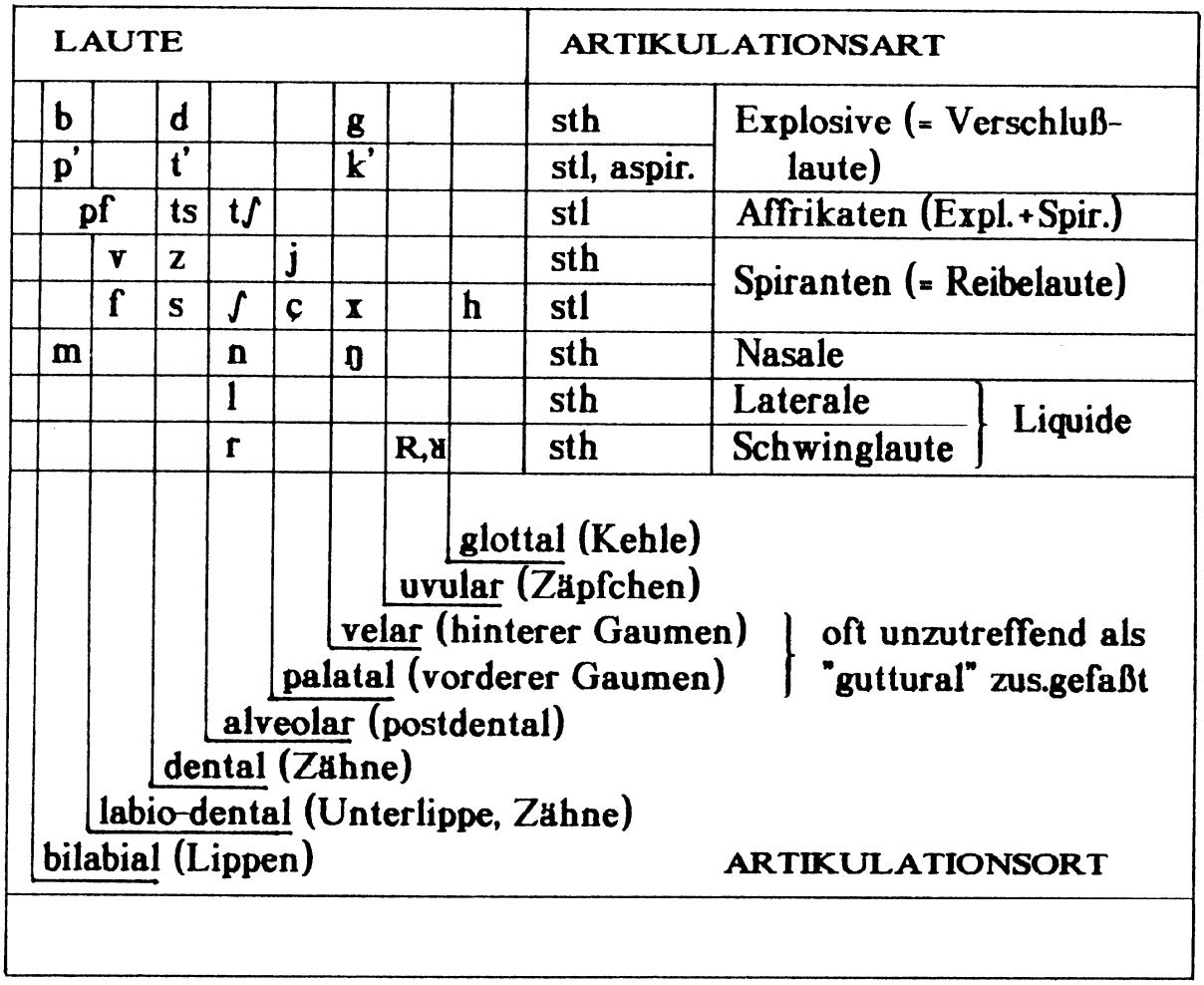


Abbildung 2: Laute, Artikulationsarten und -orte (Gross, 1990, 39).

Spätere Soundex-Varianten haben die beiden Nasale in einer Gruppe vereint. Zu diesen späteren Varianten zählt z.B. der im Rahmen des US-Census 1920 verwendete American Soundex, auch Miracode genannt, sowie der von D. Knuth 1973 veröffentlichte Simplified Soundex. Tabelle 2 zeigt die Zuordnung der Buchstaben einer Lautgruppe zum jeweiligen Simplified bzw. American Soundex-Code.

Buchstabe	Code-Ziffer
A, E, H, I, O, U, W, Y	0
B, F, P, V	1
C, G, J, K, Q, S, X, Z	2
D, T	3
L	4
M, N	5
R	6

Tabelle 2: Zuordnung von Lautgruppen zu Codes in Simplified/American Soundex.

Beim American Soundex werden die Buchstaben H und W zwischen zwei Konsonanten vorab entfernt, so dass gleichlautende Konsonanten, welche durch H oder W getrennt sind, zu einem einzelnen Code zusammengefasst werden.

Führt man Soundex beispielhaft am Namen „Johannes“ durch (Abbildung 3), so sieht man, dass im ersten Schritt der erste Buchstabe übernommen wird. Die weiteren Buchstaben werden entsprechend Tabelle 1 den jeweiligen Lautgruppen zugeordnet. Schließlich werden noch alle doppelt aufeinanderfolgenden Ziffern und alle Nullen gestrichen und der Code auf 4 Stellen gekürzt, bzw. – wie im gewählten Beispiel – durch Nullen bis auf 4 Stellen aufgefüllt. Schließlich ergibt sich der Soundex-Code J620.

```
Wort:     J O H A N N E S
Soundex:  J 0 0 0 6 6 0 2
```

Abbildung 3: Soundex am Beispiel des Wortes „Johannes“.

Die Kölner Phonetik wurde 1969 von Hans Joachim Postel entwickelt (Postel, 1969, 925-931). Das Grundprinzip der Kölner Phonetik ist dem von Soundex ähnlich, jedoch wurden die Lauttabellen besser an die deutsche Sprache angepasst. Im Gegensatz zu Soundex wird nicht jedes Zeichen isoliert betrachtet, sondern es wird ein Kontext -bestehend aus maximal einem benachbarten Zeichen- betrachtet, um den tatsächlichen Laut dieses Zeichens genauer ermitteln zu können. Ebenso gibt es für den Wortanfang (den sogenannten „Anlaut“) teilweise spezielle Kodierungen. Tabelle 3 zeigt die Zuordnung der Buchstaben einer Lautgruppe zum jeweiligen Code der Kölner Phonetik.

Buchstabe	Kontext	Code-Ziffer
A, E, I, J, O, U, Y		0
H		-
B		1
P	nicht vor H	
D, T	nicht vor C, S, Z	2
F, V, W		3
P	vor H	
G, K, Q		4
C	im Anlaut vor A, H, K, L, O, Q, R, U, X	

	vor A, H, K, O, Q, U, X außer nach S, Z	
X	nicht nach C, K, Q	48
L		5
M, N		6
R		7
S, Z		8
C	nach S, Z	
	im Anlaut außer vor A, H, K, L, O, Q, R, U, X	
	nicht vor A, H, K, O, Q, U, X	
D, T	vor C, S, Z	
X	nach C, K, Q	

Tabelle 3: Zuordnung von Lautgruppen zu Codes in der Kölner Phonetik.

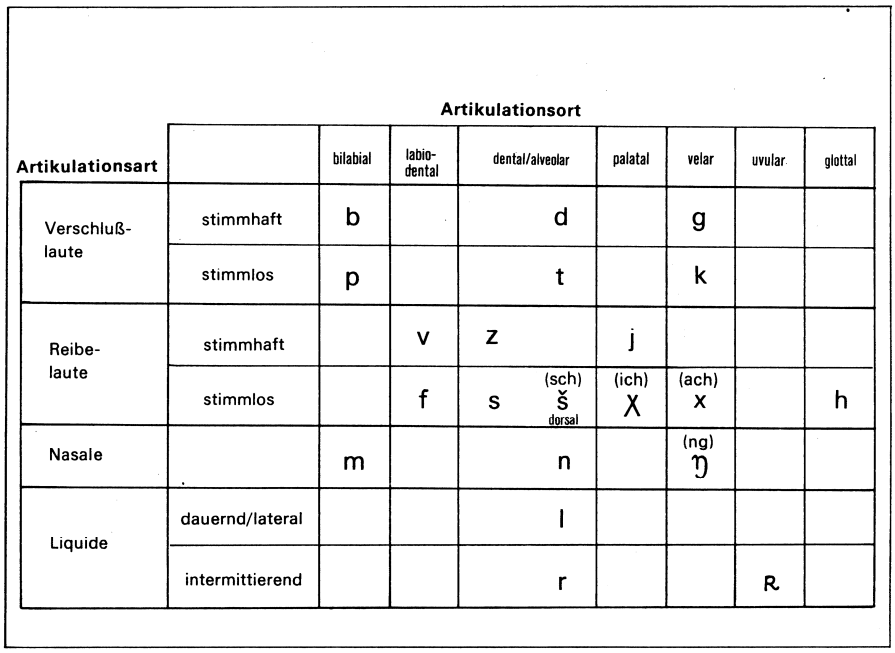
Bei der Kölner Phonetik werden die Vokale einschl. j vom glottalen h getrennt. Unter Code 1 folgen die bilabialen Verschlusslaute (b und p), unter Code 2 die dentalen Verschlusslaute (d,t). Unter Code 3 folgen die labio-dentalen Reibelaute, wobei die Kölner Phonetik den Vorteil hat, lautliche Kontexte einzubinden, beispielsweise den Effekt, dass p vor h ebenfalls zum Reibelaut (Spiranten) wird. Unter Code 4 werden die velaren Verschlusslaute zusammengefasst, wobei im Falle von c die unterschiedliche Lautung aufgrund des Kontextes berücksichtigt wird. Die Liquide l und r stehen auch hier jeweils als eigene Gruppe zur Verfügung, die Nasale werden wie beim erweiterten Soundex zusammengefasst. Die dental/alveolaren Reibelaute s,z (und in bestimmten Kontexten auch c) werden im Gegensatz zu Soundex von den velaren Verschlusslauten getrennt, was sinnvoll zu sein scheint. Es folgen unter Code 8 alle weiteren Konsonanten-Kombinationen, die zu einem dental-alveolaren Reibelaut führen.

Führt man die Kölner Phonetik am Beispiel „Johannes“ durch (Abbildung 4), so werden zunächst die Buchstaben J und O der Lautklasse 0 zugeordnet. Das H wird nicht betrachtet und auch der Vokal A wird der Lautklasse 0 zugeordnet. Entsprechende Zuordnungen werden mit den übrigen Buchstaben anhand von Tabelle 3 durchgeführt. Schließlich werden – entsprechend wie auch schon bei Soundex – alle doppelt aufeinanderfolgenden Ziffern sowie alle Nullen, die nicht am Wortanfang stehen, gestrichen. Der phonetische Code, der durch die Kölner Phonetik ermittelt wird, ist somit 068.

```
Wort:              J O H A N N E S
Kölner Phonetik:   0 ~~0~~ ~~0~~ 6 ~~6~~ ~~0~~ 8
```

Abbildung 4: Kölner Phonetik am Beispiel des Wortes „Johannes“.

Man sieht also, wie stark die Algorithmen lautliche Grundlagen wie Artikulationsart- und ort als Ordnungsprinzip für ihre Codierungen zur Anwendung bringen, wobei die Kölner Phonetik am meisten den Grundprinzipien der deutschen Sprache und deren Lautung entspricht.



Artikulationsart		Artikulationsort: bilabial	labio-dental	dental/alveolar	palatal	velar	uvular	glottal
Verschluß-laute	stimmhaft	b		d		g		
	stimmlos	p		t		k		
Reibe-laute	stimmhaft		v	z	j			
	stimmlos		f	s (sch) š dorsal	(ich) X	(ach) x		h
Nasale		m		n		(ng) ŋ		
Liquide	dauernd/lateral			l				
	intermittierend			r			R	

Abbildung 5: Konsonanteninventar des Deutschen (König, 1985, 16).

Der Grundgedanke ist bei beiden vorgestellten Algorithmen derselbe: Man möchte unterschiedliche Schreibungen eines weitgehend gleich klingenden Namens zusammenführen. Sicherlich hängt die Brauchbarkeit bzw. Anwendbarkeit eines Algorithmus stark von den abweichenden Lautsystemen einzelner Sprachen ab. Und auch das Problem der gleichen Bedeutung bei völlig anderer Lautung ist ein Problem, worauf unten noch näher einzugehen sein wird.

Metaphone (Philips, 1990, 39) wurde 1990 von Lawrence Philips als Verbesserung von Soundex entwickelt. Ebenso wie bei Soundex werden auch in Metaphone alle Vokale nach dem ersten Zeichen ignoriert und gleichlautende Zeichen zusammengefasst. Auch bei Metaphone wird wie bei der Kölner Phonetik der Kontext eines Buchstabens als wesentliche Größe zur Bildung von Lauten berücksichtigt, doch werden hier keine den phonetischen Prinzipien entsprechende Gruppen gebildet, sondern die folgenden 16 konsonantische Einzellaute zugrundegelegt, die miteinander kombiniert werden.

0BFHJKLMNPRSTWXY
(wobei X den weichen „ch“-Laut und 0 das englische „th“ repräsentiert)

Die Reduktion auf 16 Laute wird durch zahlreiche Ersetzungsregeln erreicht. Diese wiederum können um einzelsprachliche phonetische Regeln erweitert werden, so dass höhere Genauigkeit und die Erfassung verschiedener Sprachfamilien möglich wird. Die Flexibilität wird dadurch erhöht, wohingegen die Orientierung an phonetischen Systemen Einzel-Sprachen mehr entgegenkommt und innerhalb derer hohe Genauigkeit erzielen kann.

Führt man Metaphone am Beispiel „Johannes“ durch (Abbildung 6), so werden die Vokale O, A und E nicht betrachtet. Der Anfangsbuchstabe J bleibt bestehen. Beim Buchstaben H wird zunächst der Kontext betrachtet. Da auf diesen ein weiterer Vokal folgt, bleibt er bestehen. Eine Kombination des Buchstabens wie SH wäre schon vorher entdeckt worden und der entsprechenden Lautklasse zugeordnet worden. Das N bleibt wiederum bestehen, und beim S wird unterschieden, ob es der Lautklasse X oder S zuzuordnen ist (X wird am Wortanfang

durch S ersetzt). In unserem Fall trifft durch die auslautende Position Lautklasse S zu, so dass der Metaphone-Code des Namens „Johannes“ schließlich als JHNS ermittelt wird.

```
Original: J O H A N N E S
„Lemma“:  J – H – N - - S
```

Abbildung 6: Metaphone am Beispiel des Wortes „Johannes“.

Das Procedere solcher Suchverfahren bietet die Möglichkeit, in digital erschlossenen oder edierten Korpora nach Namen bzw. Personen als Träger eines Namens zu suchen, auch wenn diese nicht zuvor eigens jeweils eine Normform zugewiesen bekommen haben. Das heißt, dass man die mitunter recht schwierige Suche und Festlegung auf ein einziges Lemma, unter dem alle nur denkbaren Schreibvarianten erfasst sein müssen, um alle Namenbelege, die eine gesuchte Person bezeichnen können, zusammenzuführen, auf diese Art und Weise umgehen kann. Durch weitere Parameter der Datenbank, wie Geburtsdatum und -ort, Studienort, Studienfächer, Abschlussjahr, Abschlussgrad etc. kann eine Fokussierung auf eine gesuchte Person erreicht werden oder deren akademischer Werdegang unter Umständen nachvollzogen werden. Dies wäre für die prosopographische Erforschung des hohen und späten Mittelalters sowie der frühen Neuzeit eine enorme Erleichterung und trägt den größeren Datenmengen Rechnung, die im Gegensatz zum Frühmittelalter bestehen und eine handverlesene Lemmatisierung unmöglich machen.

Eine Herausforderung stellt dabei nur die große Zeitspanne und, mit Blick auf die Netzwerke, welche die mittelalterlichen europäischen Universitäten erzeugten, große geographische Ausdehnung dar, die zu vielen sprachlichen Variationen eines Namens oder gar zu dessen kompletter Übersetzung führen können (wie bei unserem Beispiel-Fall Johannes, welcher zu Giovanni, Juan, Jean, John, Jan, Jens, János, Iwan oder Seán bzw. Siobhan als weiblicher Variante werden kann).

In einem tabellarischen Vergleich der Verfahren für verschiedene Namenspaare (Tabelle 4) kann man sehen, dass Metaphone und Soundex einen starken Fokus auf identische Wortanfänge legen. So erkennen diese – im Gegensatz zur Kölner Phonetik – nicht die Lautgleichheit der Namen *Johannes* und *Ioannes*.

Name	Soundex	Kölner Phonetik	Metaphone
Johannes	J620	068	JHNS
Ioannes	I620	068	INS
Giovanni	G160	436	JFN
Juan	J600	06	JN
Jean	J600	06	JN
John	J600	06	JN
Jan	J600	06	JN
Jens	J620	068	JNS
Janos	J620	068	JNS
Iwan	I600	036	IWN
Sean	S600	86	SN
Shiobhan	S160	816	XBHN

Tabelle 4: Vergleich der Verfahren für Varianten/Übersetzungen des Namens *Johannes*.

Folglich wären Möglichkeiten zu diskutieren, einzelne, auf bestimmte Sprachgruppen sehr gut spezifizierte Suchprogramme (bspw. Kölner Phonetik) miteinander so zu verknüpfen, dass nicht ein Programm mehr oder weniger gut Germanisch, Romanisch und Slawisch komplett beherrschen muss, sondern ein Meta-Modul Namensucheingaben wie bspw. „Johannes“ im speziell-germanischen Modul (vielleicht sogar in Nieder- und Hochdeutsch sowie Angelsächsisch geschieden) sucht und zugleich unter Varianten von „Giovanni“ im speziell-romanischen Modul sowie unter „János“ o.ä. im speziell-slawischen Modul nachschaut. Dies wäre insofern sinnvoll, als die Studierenden in dt. Hochschulen durchaus auch aus anderen europäischen Sprachgebieten kamen.

Auch die humanistischen Namen und Latinisierungen kommen hier ins Spiel, denn auch sie können ein und denselben Namen in einem gänzlich anderen sprachlichen Gewand erscheinen lassen. Dazu lässt sich ein Heidelberger Beispiel anführen: Der schweizerische Humanist Johannes Sapidus (oder Witz) latinisierte die Namen einiger seiner Schüler, weil er sie für zu ungebildet (*bara nomina*) hielt. Ein niederländischer Theologiestudent namens Puts nannte sich bereits Puteanus, jedoch erklärte ihm Prof. Zacharias Ursinus, man kenne hier im Süden keine putten (Brunnen), sondern nur Fontänen, weshalb er fortan Fontanus heißen solle (Rentenaar, 2002, 163).

Name	Soundex	Kölner Phonetik	Metaphone
Puts	P320	18	PTS
Puteanus	F352	1268	PTNS

Tabelle 5: Vergleich der Verfahren für *Puts* und *Puteanus*.

Mit Blick auf Namenspaare wie Gottfried/Godofridus und Bocholt/Buchholdius kann man sehen, dass selbst auf den ersten Blick im Schriftbild sehr unterschiedliche Namenspaare von allen drei Verfahren als lautgleich erkannt werden. Hier zeigt sich auch eine mögliche Intention der Kürzung durch Soundex auf 4 Zeichen: durch diese Kürzung werden zusätzliche Endungen (z.B. us) ausgeblendet, so dass diese nicht zu verschiedenen Codes führen. Bei entsprechend langen Namen, die sich vor allem im hinteren Teil unterscheiden, kann dies natürlich zu Fehlentscheidungen führen.

Name	Soundex	Kölner Phonetik	Metaphone
Gottfried	G317	42372	KTFRT
Godofridus	G317	423728	KTFRTS
de/von Bocholt	B243	1452	BXLT
Buchholdius	B243	14528	BXLTS

Tabelle 6: Vergleich der Verfahren für *Gottfried/Godofridus* und *de Bocholt/Bucholdius*.

An den letzten beiden Namenspaaren Weber/Textor sowie Fischer/Piscator kann man sehen, dass durch solche Verfahren nicht alle Entwicklungen erkannt werden können. Lediglich, wenn die Namen lautgleich bleiben und sich nur die Schreibart unterscheidet, können diese Namen als verwandt identifiziert werden. Ändert sich ein Name z.B. aufgrund von Latinisierung im Lautbild, wobei die Bedeutung des Namens (mehr oder weniger) gleich bleibt, so kann die Verwandtschaft nicht durch diese Verfahren entdeckt werden - hierfür sind weitergreifende Ansätze mit einem hinterlegten Lexikon erforderlich.

Name	Soundex	Kölner Phonetik	Metaphone
Textor	T237	24826	TKSTR
Weber	W160	317	WBR
Piscator	P237	1827	PSKTR
Fischer	F260	387	FSXR

Tabelle 7: Vergleich der Verfahren für *Textor/Weber* und *Piscator/Fischer*.

Nicht zuletzt spielt der sprachliche Hintergrund dessen, der bspw. die Namen Studierender in Matrikellisten eintrug eine Rolle, wenn er mit dem Betroffenen nicht identisch war, dieser sich also nicht selbst eigenhändig eingeschrieben hat. In diesem Fall wiederum können paläographische Abgleiche zusätzlich für Eingrenzungsmöglichkeiten sorgen, was bei der Konzeption der Datenbank, in welche die gescannten Originalseiten implementiert sein sollen, zu erweiterten Suchmöglichkeiten führt.

Abbildung 7: Web-Datenbank der Universitätsmatrikeln, Bereich „Namenbelege“.

Sicherlich werden die Suchmotoren auf phonetischer Grundlage lernen müssen und sich den genannten Problemen stellen müssen. Aber gerade in diesen Problemen liegt auch der große Reiz und die große Chance, auf historischem, namenkundlichem sowie auf datentechnischem Fachgebiet vertiefte Forschungen anzustellen und Lösungsansätze zu entwickeln, die auch auf andere Bereiche übertragen werden können, bei denen es um große Namenmengen und -varianten geht.

Literatur

Förstemann, Ernst. 1900. *Altdeutsches Namenbuch, Erster Band: Personennamen*. 2. völlig umgearbeitete Aufl. Bonn.

Geuenich, Dieter. 2002. Zur Entstehung und Entwicklung der Familiennamen im hohen Mittelalter. In: Dieter Kremer (ed.), *Onomastik. Akten des 18. Internationalen Kongresses*

für Namenforschung, Trier 12.-17. April 1993, Band 6: Namenforschung und Geschichtswissenschaften. Literarische Onomastik, Namenrecht, Ausgewählte Beiträge (Ann Arbor, 1981), Tübingen, 41-48.

Geuenich, Dieter; Hawicks, Heike. 2008. Probleme der Lemmatisierung frühmittelalterlicher Personennamen im interdisziplinären Projekt ‚Nomen et gens'. In: Eva Brylla / Mats Wahlberg (ed.), *Proceedings of the 21st International Congress of Onomastic Sciences (ICOS), 19-24 August 2002 Uppsala*, Volume 4. Uppsala, 81-90.

Geuenich, Dieter; Runde, Ingo (ed.). 2006. *Name und Gesellschaft im Frühmittelalter. Personennamen als Indikatoren für sprachliche, ethnische, soziale und kulturelle Gruppenzugehörigkeiten ihrer Träger*. Deutsche Namenforschung auf sprachgeschichtlicher Grundlage 2. Hildesheim / Zürich / New York.

Gross, Harro. 1990. *Einführung in die germanistische Linguistik*. München.

Kaufmann, Henning. 1968. *Ergänzungsband zu Ernst Förstemann Altdeutsche Personennamen*. München / Hildesheim.

Knuth, Donald E. 1973. *The Art of Computer Programming*, Volume 3: Sorting and Searching, Reading, Mass.

König, Werner. 1985. *dtv-Atlas zur deutschen Sprache. Tafeln und Texte*. 6. Aufl. München.

Koß, Gerhard. 2002. *Namenforschung. Eine Einführung in die Onomastik*. 3. Aufl. Tübingen.

Mitterauer, Michael. 1993. *Ahnen und Heilige. Namengebung in der europäischen Geschichte*. München.

Philips, Lawrence. 1990. Hanging on the Metaphone. *Computer Language* 7, No. 12, Dec. 1990, 39.

Postel, Hans Joachim. 1969. Die Kölner Phonetik. Ein Verfahren zur Identifizierung von Personennamen auf der Grundlage der Gestaltanalyse. *IBM-Nachrichten*, 925-931.

Rentenaar, Rob. 2002. Humanismus und Familiennamen. Zur Entstehung und Verbreitung der humanistischen Familiennamen in Nordwesteuropa. In: Dieter Kremer (ed.), *Onomastik. Akten des 18. Internationalen Kongresses für Namenforschung, Trier 12.-17. April 1993, Band 6: Namenforschung und Geschichtswissenschaften. Literarische Onomastik, Namenrecht, Ausgewählte Beiträge (Ann Arbor, 1981)*, Tübingen, 161-167.

Runde, Ingo; Voigt, Tobias. 2002. Neue Möglichkeiten der EDV bei der Erforschung mittelalterlicher Personennamen. In: Ana Isabel Boullón Agrelo (ed.), *Actas do XX congreso internacional de ciencias onomásticas, Santiago de Compostela, 20-25 setembro 1999*. A Coruña, 1249-1264.

Russell, Robert C. 1918. *US-Patent 1, 261, 167, Apr. 02, 1918*.

Stefan Böttcher. Universität Paderborn
stb@uni-paderborn.de

Rita Hartel. Universität Paderborn
rst@upb.de

Heike Hawicks. Universität Duisburg-Essen
heike.hawicks@uni-due.de
Deutschland