

Ruth Christmann, Vera Hildenbrandt, Thomas Schares

Ein „heiligthum der sprache“¹ digitalisiert:

Das Deutsche Wörterbuch von Jacob und Wilhelm Grimm
auf CD-ROM und im Internet

Dem von Jacob und Wilhelm Grimm begründeten Deutschen Wörterbuch² wird in der deutschen Wissenschaftsgeschichte zu Recht eine Sonderstellung eingeräumt, handelt es sich doch um das deutschsprachige Wörterbuch mit der längsten Bearbeitungszeit und der reichhaltigsten Dokumentation des Deutschen.

Seit dem 01.11.1998 läuft an der Universität Trier ein DFG-Projekt zur Retrodigitalisierung dieses Wörterbuchs³; wie dabei vorgegangen wird, erläutert der folgende Beitrag. Ein kurzer geschichtlicher Abriss soll jedoch zunächst die Bedeutung des DWB für die Forschung darstellen und den lexikographiegeschichtlichen Hintergrund des Projektes skizzieren, bevor dann der Einsatz und die Rolle von TUSTEP bei der Retrodigitalisierung eines so umfassenden und – wie seine Geschichte zeigt – nicht unproblematischen Werkes erörtert werden. Im Anschluss an die Beschreibung des Wegs von der Dateneingabe zur elektronischen Publikation wird ein weiterer Schwerpunkt auf den Problemen und Schwierigkeiten liegen, welche die Fehlersuche und Sonderzeichenbehandlung in der TUSTEP-konformen Datengrundlage aufwerfen.

1. Kurzgefasste Geschichte des DWB

„Was ist eines wörterbuchs zweck? [...] Es soll ein heiligthum der sprache gründen, ihren ganzen schatz bewahren, allen zu ihm den eingang offenhalten.“⁴

Der Begründung eines solchen, von ihm so genannten „heiligthum[s] der sprache“ widmet Jacob Grimm die letzten 25 Jahre seines Lebens, nachdem er und sein Bruder Wilhelm im Jahre 1838 mit dem Projekt eines großen historisch-entwicklungsbezogenen Wörterbuchs des neuhochdeutschen Wortschatzes von etwa 1450 bis zur Bearbeitungsgegenwart betraut wurden.

Da sich von Anfang an das Bestreben der Brüder darauf richtet, „eine weit vollere und lebendigere sammlung aller deutschen wörter vorzunehmen als sie noch stattgefunden hat“⁵, erfolgt in den darauf folgenden Jahren zunächst eine breit an-

1 JACOB GRIMM: Vorrede zum ersten Band des DWB, S. XII.

2 Wenn hier und im Folgenden vom Deutschen Wörterbuch, Grimm oder DWB gesprochen wird, so bezieht sich dies stets auf die 1961 abgeschlossene Ausgabe des Wörterbuchs.

3 Das Projekt wird gefördert im Rahmen des DFG-Programms „Retrospektive Digitalisierung von Bibliotheksbeständen“, vgl. http://www.SUB.uni-goettingen.de/ebene_2/2_vdfpro.htm.

4 JACOB GRIMM: Vorrede zum ersten Band des DWB, S. XII.

5 Ebd., S. IV.

gelegte Exzerption überwiegend literarischer Quellen. Mehr als 80 Mitarbeiter überall im deutschsprachigen Raum sammeln ca. 600 000 Belege für den Wortschatz von A bis Z und schaffen somit eine in der Geschichte der deutschen Lexikographie ganz neue Materialbasis, die sich jedoch in der Folgezeit immer wieder als zu schmal erweist und stetiges Nachexzerpieren erfordert. Erst 1849, elf Jahre nach der ersten Ankündigung des DWB in der *Leipziger Allgemeinen Zeitung*, beginnen Jacob und Wilhelm Grimm mit der redaktionellen Ausarbeitung des Wörterbuchs. 1852 erscheint zur Leipziger Frühjahrsmesse die erste Lieferung [A – Allverein], 1854 der erste, von Jacob Grimm verfasste Band [A – Biermolke] mit seiner bemerkenswerten, im Druck 68 Spalten umfassenden Vorrede, in der er aus der Wörterbuchpraxis heraus Rechenschaft gibt über Gegenstand, Aufgabe und Anlage des DWB und außerdem seine Wörterbuchtheorie darlegt.⁶

Demnach soll das DWB auf der Basis historischer Sprachforschung die Entwicklung des Deutschen erschließen. Es soll den neuhochdeutschen schriftsprachlichen Wortschatz von der Mitte des 15. Jahrhunderts bis zur Bearbeitungsgegenwart darstellen. Dazu gehören auch umgangssprachliche, derbe, anstößige Wörter, älterer Sonder- und Fachwortschatz zum Beispiel aus Berg- und Ackerbauschriften, Kriegs-, Koch- und Arzneibüchern etc. und – soweit sein Sprachgebrauch im Hochdeutschen nachweisbar ist – regional begrenztes und mundartliches Wortgut. Weitgehend ausgeschlossen aus dem Stichwortbestand werden dagegen reine Mundartwörter, niederdeutsches, dem Niederdeutschen entlehntes und formal nicht assimiliertes Fremdwortgut und Eigennamen. Den „mächtigsten und gewaltigsten zeugen der sprache“⁷ entnommene Belege sollen die Bedeutungen und den Gebrauch der in streng alphabetischer Ordnung aufgenommenen Wörter dokumentieren und „ihre ganze geschichte vortragen“.⁸

Gewünschte Adressaten dieses historisch ausgerichteten Belegwörterbuchs des Deutschen sind für Jacob Grimm nicht punktuell nach Informationen suchende Benutzer, sondern „leser jedes standes und alters“ [Hervorhebung d. Verf.]⁹, das heißt, das DWB soll nicht primär ein Nachschlagewerk, sondern auch und vor allem ein Lese- und Lehrbuch werden, in dem nach Möglichkeit das ganze Volk „zum hausbedarf, und mit verlangen, oft mit andacht“¹⁰ liest. Mit anderen Worten: Jacob Grimm strebt eine weite Verbreitung des Wörterbuchs an und erhofft sich davon sowohl eine spracherzieherische als auch eine politische Wirkung: Das DWB soll einen Beitrag zur Regeneration der von den Grimms als verarmt empfundenen Gegenwartssprache leisten, das heißt Fehlentwicklungen in der gegenwärtigen Sprache und im Sprachgebrauch entgegensteuern, indem es seinen Lesern eine tie-

6 Eine ausführliche Darstellung von Jacob Grimms Wörterbuchtheorie gibt DÜCKERT (1987a, S. 8a–21a).

7 JACOB GRIMM: Vorrede zum ersten Band des DWB, S. XXV.

8 Ebd., S. XXXVIII.

9 Ebd., S. XII.

10 Ebd., S. XIII.

fere Einsicht in die deutsche Sprache, ihre Geschichte und ihre Literatur vermittelt und dadurch ihr Sprachgefühl und ihr sprachliches Selbstbewusstsein stärkt. Auf diese Weise soll es ihnen zu einer eigenen sprachlichen und – da sie sich durch die gemeinsame Sprache als Angehörige eines Volkes ausweisen – zugleich auch nationalen Identität verhelfen.¹¹

In der Wörterbuchpraxis Jacob Grimms haben diese theoretischen Überlegungen Auswirkungen auf Quellen- und Stichwortauswahl und finden Umsetzung in einem ihm ganz eigenen Wörterbuchstil: Ziel Jacob Grimms ist es, die aufgenommenen Wörter als Resultat historischer Prozesse zu erklären. Auf dem Weg zu diesem Ziel folgt er dem etymologischen Prinzip der Wortklärung. Er setzt beim neuhochdeutschen Wortschatz an, führt die Wörter über vergleichbare Formen der germanischen und indogermanischen Sprachen zurück und ermittelt nach Möglichkeit ihren Ursprung und von dort aus ihre formalen und begrifflichen Grundlagen. Im Rückgriff auf diese in der Etymologie verankerte geschichtliche Basis führt er auf neuhochdeutscher Ebene die Erklärung von Wortgestalt, Bedeutungsentwicklung und Wortgebrauch vor, der durch Belege dokumentiert wird.¹² Dabei beschränkt er sich nicht auf die Darstellung von Fakten, sondern läßt auch eigene Überlegungen und Spekulationen einfließen. Häufig werden der etymologische Teil und die Erklärungen zu Bedeutung und Gebrauch stark miteinander verwoben, nicht selten werden sogar artikelübergreifende Zusammenhänge hergestellt. Auf diese Weise hat Jacob Grimms Wörterbuchstil erörternden und abhandlungsartigen Charakter, werden seine Wörterbuchartikel quasi zu einem wissenschaftlichen Diskurs, in den er den Benutzer/Leser einbezieht.¹³

Wilhelm Grimm geht andere Wege als sein Bruder: Sein Schwerpunkt liegt auf der Wortsemantik und nicht so sehr auf der Etymologie, die er im Prinzip nur dann gibt, wenn sie als gesichert gelten kann. Wie denn überhaupt das DWB für ihn anders als für Jacob weniger ein Ort der Sprachforschung ist, das heißt, im Vordergrund steht nicht das forschende, erörternde, spekulierende Suchen nach Ergebnissen, sondern die auf Fakten beschränkte, methodischere Darstellung der Ergebnisse.¹⁴

Deutlich unterscheidet sich so die Wörterbuchpraxis Wilhelms¹⁵ von der Jacobs¹⁶, so dass schon für die von ihnen geprägte erste Bearbeitungsphase des Deutschen Wörterbuchs (1838 bis 1863), in der weniger als 9 % des Gesamtwörterbuchs erarbeitet werden, nicht von einer einheitlichen Wörterbuchgestaltung gesprochen

11 Vgl. zu diesem sogenannten 'nationalpädagogischen' Programm Jacob Grimms vor allem BÄHR (1984b, S. 392–396) und ders. (1991, S. 5–8).

12 Vgl. BÄHR (1984b, S. 401/402) und ders. (1991, S. 19/20).

13 Zu Jacob Grimms eigenwilligem Wörterbuchstil vgl. PÜSCHEL (1991, S. 67–87).

14 Zur Wörterbucharbeit Wilhelm Grimms vgl. vor allem BÄHR (1991, S. 20–22), DÜCKERT (1987, S. 37b–44b) und PÜSCHEL (1991, S. 87–90).

15 Von Wilhelm bearbeiteter Abschnitt: Buchstabe D.

16 Von Jacob bearbeiteter Abschnitt: Buchstaben A, B, C, E, F.

werden kann. Eine wirklich einschneidende Zäsur in Zielsetzung und Bearbeitung des Wörterbuchs stellt allerdings erst der Tod Jacob Grimms 1863 dar.

Der Übergang zur zweiten Bearbeitungsphase des DWB (1863 bis 1908)¹⁷ geht einher mit einem Wechsel vom sprachwissenschaftlich zum philologisch orientierten Wörterbuch.¹⁸ Die deutsche Philologie, von deren „gedeihen und [...] wirkung“ das „wiedererstehen der [deutschen] nation“¹⁹ ganz wesentlich abhängig gemacht wird, übernimmt das Wörterbuch als nationales Werk, das „bewusstsein und gefühl der eignen deutschen art“²⁰ wecken soll. Aus dieser Umorientierung ergeben sich grundlegende Änderungen: Zum einen wird der Hausbuchgedanke Jacob Grimms und damit seine sprachpädagogische Absicht abgelöst von der Idee des DWB als einem Thesaurus der ganzen deutschen Sprache.²¹ Zum anderen wird Jacob Grimms etymologisches Prinzip der Wortgeschichtsbeschreibung, das schon von Wilhelm Grimm zurückhaltender gehandhabt wurde, von den ersten Fortsetzern des Wörterbuchs – unter ihnen Rudolf Hildebrand, Karl Weigand, Moriz Heyne, Hermann Wunderlich und Matthias Lexer – relativiert. Die strenge Etymologisierung der Bedeutungsgeschichte wird aufgegeben, das heißt, die Bedeutungsgeschichte wird aus der Etymologie gelöst, und neben der breiten Behandlung von Etymologie und Formgeschichte erfahren jetzt die Darstellung des Wortgebrauchs und seine Interpretation eine stärkere Berücksichtigung. Da es allerdings weder eine verbindliche Wörterbuchkonzeption noch eine einheitliche Redaktion gibt und die Artikelautoren eigenverantwortlich und ohne wechselseitige Absprache arbeiten, sind ihre Blickrichtungen, Ausgangspunkte, methodischen Ansätze und Schwerpunkte unterschiedlich, so dass es zur Ausbildung verschiedener, bisweilen eigenwilliger Darstellungsweisen kommt. Um die Jahrhundertwende nehmen Uneinheitlichkeiten und – bedingt durch das ausufernde Streben nach innerer und äußerer Vollständigkeit – die Breite der Darstellung über Gebühr zu, das Arbeitstempo verlangsamt sich erheblich, nicht zuletzt auch durch die unzureichende Materialgrundlage.

Dem versucht eine erste Reorganisation des DWB entgegenzuwirken, die die dritte Bearbeitungsphase des Deutschen Wörterbuchs (1908 bis 1930) einleitet.²² Im Jahr 1908 – bis zu diesem Zeitpunkt sind etwas mehr als 50 % des Wörterbuchs erarbeitet – übernimmt die Deutsche Kommission der Königlich-Preußischen Akademie der Wissenschaften zu Berlin die wissenschaftliche Leitung des DWB. Mit Hilfe organisatorischer Maßnahmen versucht man die in der zweiten Arbeitsphase offenbar gewordenen Mängel und Schwierigkeiten zu beheben und so den Fortgang des Unternehmens zu sichern. Wichtigste Neuerungen sind die Gründung einer Zentralsammelstelle für die Beschaffung von Belegmaterial in Göttingen, die in

17 Vgl. zu dieser Bearbeitungsphase des DWB insbesondere HUBER (1987) und SCHRÖTER (1987).

18 Vgl. BAHR (1991, S. 22).

19 RUDOLF HILDEBRAND: Vorrede zum fünften Band des DWB, S. I.

20 Ebd., S. I.

21 Vgl. BAHR (1984a, S. 497a).

22 Vgl. zu dieser Bearbeitungsphase des DWB insbesondere BRAUN (1987, S. 126b–131b).

kurzer Zeit mehr als zwei Millionen Belege für die noch ausstehenden Teile des Wörterbuchs exzerpiert, und die Erhöhung der Mitarbeiterzahl. Außerdem ist mit dem Einsetzen eines akademischen Leiters, der eine Fahnenkorrektur jedes Artikels lesen soll, erstmals die Möglichkeit unmittelbarer Einflussnahme auf die Artikelarbeit gegeben. Doch auch diese ersten Ansätze zu einer zentralen Redaktion und der Festsetzung eines Normalmaßes für den Artikelumfang können eine zunehmende methodische Desorientierung in der Wörterbuchpraxis nicht verhindern. Die Arbeitsverfahren und Artikelstrukturen sind in dieser Phase sehr heterogen.

Zwar bemüht sich Arthur Hübner um eine Neuorientierung in der Wörterbuchpraxis, indem er versucht, durch die Artikelgliederung die inhaltliche und geschichtliche Struktur des Wortes wiederzugeben und dadurch eine konzentriertere und durchsichtigere Anlage der Wörterbuchartikel zu erreichen.²³ Doch trotz der organisatorischen Neuerungen und Hübners angestrebter Instrumentalisierung der Artikelgliederung wird in dieser dritten Phase im großen und ganzen das Verfahren des vorangehenden Arbeitsabschnitts beibehalten und die Effektivität der praktischen Arbeit nicht gesteigert. Zwischen 1908 und 1930 erscheinen nur drei Bände des Wörterbuchs.

Daher entschließt man sich zu einer zweiten Reorganisation des DWB, die den vierten Arbeitsabschnitt in der Geschichte des Wörterbuchs (1930 bis 1961)²⁴ einleitet. Im Jahr 1930 wird in Berlin unter der Leitung Arthur Hübners eine ständige Arbeitsstelle eingerichtet, in der eine Gruppe von hauptberuflich tätigen Mitarbeitern unter redaktioneller Leitung ihrer Wörterbucharbeit nachgehen kann. Verbunden mit der Einrichtung dieser Arbeitsstelle ist der Aufbau einer allen Mitarbeitern zur Verfügung stehenden Spezialbibliothek und die Verlegung des Wortarchivs von Göttingen nach Berlin im Jahr 1934, wodurch erstmals die Voraussetzungen einer konzentrierten Wörterbucharbeit gegeben sind. Des weiteren werden 1930/31 die besten Erfahrungen aus den vorhergehenden Arbeitsphasen des DWB, vor allem die Hübners, in die schriftliche Form von Arbeitsrichtlinien umgesetzt. Diese Richtlinien steuern bis zum Abschluss des DWB die praktischen Arbeiten; sie legen die Zielsetzungen des Wörterbuchs fest und regeln auf formaler und inhaltlicher Ebene Anlage und Aufbau der Wörterbuchartikel.

In der Praxis führen die Anweisungen zu einer Straffung des Formteils, der gegenüber dem Bedeutungsteil jetzt mehr die Funktion einer Einleitung erhält. Im Zentrum des Artikels stehen Bedeutungen und ihre geschichtlichen und inhaltlichen Zusammenhänge.²⁵ Um größtmögliche Einheitlichkeit in der technischen Darstellungsweise zu erreichen, das Einfließen persönlicher Vorlieben der Artikelautoren weitgehend zu unterbinden und ein Ausufern der Darstellung zu verhindern, wacht die Leitung der Arbeitsstelle zunehmend darüber, dass alle Beiträge einer angemessenen Redaktion unterzogen werden.²⁶

23 Die Bemühungen Hübners werden ausführlich dargestellt von BAHR (1991, S. 31–36).

24 Vgl. zu dieser Bearbeitungsphase vor allem BRAUN (1987, S. 132a–150b).

25 Vgl. BAHR (1984a, S. 497a).

Zwar bewährt sich diese erneute Reform der Arbeitsorganisation insgesamt, und es kann ein relativ einheitlicher Arbeitsstil durchgesetzt werden, dennoch werden auch in dieser letzten Arbeitsphase bei genauerer Betrachtung Unterschiede zwischen den Artikeln sichtbar, die nicht zuletzt daher rühren, dass einige der freien Mitarbeiter sich der Autorität der zentralen Leitung entziehen und der alten Darstellungsweise verhaftet bleiben. Immerhin aber kann die jährliche Leistung deutlich gesteigert werden, und im Januar 1961 erscheint schließlich die letzte der 380 Lieferungen des DWB. 1971 folgt das Quellenverzeichnis, das über 25 000 Titel und Verweise der im DWB systematisch oder nur gelegentlich benutzten Quellen umfasst.

Was läßt sich nun nach diesem knapp skizzierten Abriss der Geschichte des Deutschen Wörterbuchs zusammenfassend über dieses riesige Werk sagen? Was ist aus den Planungen und Konzepten der Brüder Grimm geworden?

Ursprünglich hatte Jacob Grimm für das DWB einen Umfang von etwa sechs bis sieben Bänden und eine Bearbeitungszeit von sechs bis zehn Jahren vorgesehen.²⁷ Tatsächlich erschienen 16 Bände in 32 Teilbänden mit insgesamt 67 744 Spalten und etwa 350 000 Stichwörtern in mehr als einhundert Jahren. Ebenso wie Jacob Grimm den Umfang des Wörterbuchs unterschätzt hat, hat er zweifellos die Größe des in Frage kommenden Leserkreises und den Einfluss des Werkes auf das Volk überschätzt: Seine Hoffnung auf eine Breitenwirkung des DWB erfüllte sich nicht. Nicht „das ganze Volk“, sondern Sprach- und Literaturwissenschaftler, Philologen und andere an sprachhistorischen Informationen interessierte Fachwissenschaftler bildeten und bilden seinen überwiegenden Benutzerkreis. Damit musste zwangsläufig auch das (ohnehin von seinen Nachfolgern nicht fortgesetzte) nationalpädagogische Programm Jacob Grimms fehlschlagen. Gewahrt dagegen blieb über alle Bearbeitungsphasen hinweg die Intention einer historischen Beschreibung des neuhochdeutschen Wortschatzes.²⁸ Somit hat sich Jacob Grimms ursprüngliches Programm, wie er es vor allem in seiner Vorrede zum ersten Band des DWB dargelegt hat,²⁹ nur als Gerüst bewährt, das es späteren Bearbeitern erlaubte, eigenen Neigungen nachzugehen und eigene Schwerpunkte zu setzen.³⁰

Durch das Fehlen einer für alle Bände grundlegenden Wörterbuchkonzeption, die überlange Bearbeitungsgeschichte und die Tatsache, dass das DWB ein Werk vieler ist, durch das Einfließen sowohl zeitgenössischer Vorstellungen und Vorlieben der jeweiligen Bearbeiter als auch jeweils aktueller Erkenntnisse aus Sprachwissenschaft, Philologie und Geschichtswissenschaft finden immer neue Wörterbuchstile Eingang in die ständig erweiterte und differenzierte Darstellung, entstehen Wörterbuchartikel, die in vielerlei Hinsicht nicht den Erwartungen entsprechen, die heu-

26 Vgl. BRAUN (1987, S. 134b–135b).

27 Vgl. DÜCKERT (1987b, S. 40).

28 Vgl. BAHR (1984a, S. 494a/b).

29 Vgl. oben.

30 Vgl. HORLITZ (1991, S. 414).

tige Benutzer gewöhnlich an Wörterbuchartikel stellen. Das heißt, dass sich der heute im Grimm Nachschlagende mit hohen Anforderungen des Wörterbuchs an seine Person konfrontiert sieht: Er muss nicht nur auf die Uneinheitlichkeit des DWB und die unterschiedliche Quantität und Qualität der Artikel gefasst sein, er muss sich auch darauf einstellen, dass sich die bereitgestellten Informationen sehr oft nicht durch punktuelles Nachschlagen entnehmen lassen, sondern nur durch Lesen des gesamten Artikels.³¹ Somit wird das DWB zu Recht als „problematisches Monumentalwerk“³² bezeichnet.³³

Gleichzeitig darf aber nicht außer Acht bleiben, dass mit dem Deutschen Wörterbuch nicht nur die umfang- und materialreichste Dokumentation des Deutschen gegeben ist, sondern auch gerade aufgrund seiner langen Geschichte und der wechselnden lexikographischen Praxis ein unvergleichliches Zeugnis der deutschen Wissenschaftsgeschichte des 19. und 20. Jahrhunderts und somit ein für Lehre und Forschung unerlässliches Grundlagenwerk.

Die systematische elektronische Aufbereitung soll dieses Auskunftsmittel und Forschungsinstrument nicht nur einem größeren Benutzerkreis zugänglich machen, also, um es mit den Worten Jacob Grimms zu sagen, allen zu diesem „heiligthum der sprache [...] den eingang offenhalten“³⁴, sondern es auch mit erweiterten Benutzungsmöglichkeiten versehen, Möglichkeiten, die über die Probleme bei seiner Benutzung hinweghelfen sollen.

2. Von der Dateneingabe zur elektronischen Publikation

Im Folgenden wird gezeigt, wie die Probleme bei der Retrodigitalisierung des Deutschen Wörterbuchs der Brüder Grimm, denen sich der Benutzer (und Digitalisierer) aufgrund des Alters, des Umfangs und der uneinheitlichen Struktur des DWB stellen muss und die sich bei anderen Wörterbüchern nicht oder nur in geringem Umfang ergeben, in Angriff genommen werden. Dabei wird insbesondere auf die Rolle von TUSTEP in der Projektarbeit eingegangen.

Eine Seite des DWB enthält etwa 9 000 Zeichen, das gesamte Deutsche Wörterbuch 35 000 Seiten. Zwanzig Spalten des Wörterbuchttexts im TUSTEP-Format entsprechen rund 120 KB. Einschließlich der TUSTEP-Kodierungen ergibt das etwa 11 MB pro Band und, hochgerechnet auf den gesamten „Grimm“, etwa 360 bis 370 MB, eine Zahl, die sich durch die fortschreitende Auszeichnung mit SGML-Tags um ein Vielfaches erhöhen wird. Die Datensicherheit spielt bei diesen Datenmengen eine große Rolle, deshalb erfolgt der Großteil der Arbeiten auf einem UNIX-Großrechner im Rechenzentrum der Universität Trier.

31 Zu Problemen in der Benutzung des DWB vgl. vor allem HORLITZ (1991).

32 KIRKNESS/KÜHN/WIEGAND (1991, S. VIII).

33 Vgl. dazu auch die folgenden Beiträge, die sich anhand einzelner Artikelanalysen mit der problematischen Konzeption des Wörterbuchs beschäftigen: BERGMANN (1999, S. 339–345), SCHLAEFER (1999, S. 204–207), SCHULZ (1999, S. 56–60).

34 JACOB GRIMM: Vorrede, S. XII.

Für die Projektarbeit hält TUSTEP einige wertvolle Funktionen bereit: Zum einen können mehrere sogenannte 'Sitzungen' bereitgestellt werden, die als eigene 'Arbeitsplätze' eingerichtet werden, um verschiedene Arbeitsschritte weitgehend unabhängig voneinander durchzuführen. Zum anderen wurde für die Version TUSTEP 2000 ein Kontrollmechanismus implementiert, der verhindert, dass gleichzeitig auf ein- und dieselbe Datei schreibend zugegriffen wird.

Die genaue Vorgehensweise des Projekts stellt sich wie folgt dar:³⁵

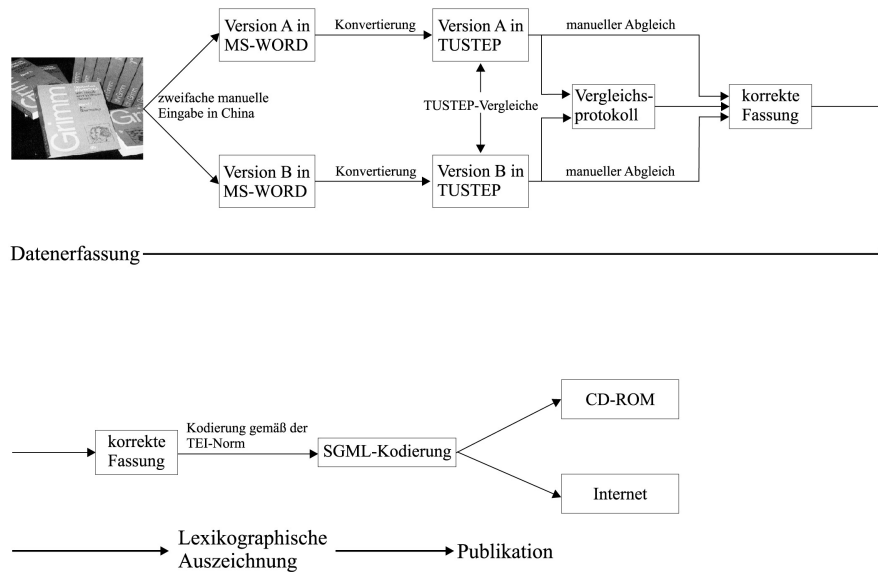


Abb. 1: Die Arbeitsschritte von der Eingabe des DWB bis zur Internet-/CD-ROM-Version.

Die Dateneingabe erfolgt durch eine chinesische Firma, die bereits mit der Datenerfassung für das Projekt „Mittelhochdeutsche Wörterbücher auf CD-ROM und im Internet“, das im Beitrag von JOHANNES FOURNIER in diesem Band vorgestellt wird, betraut war.³⁶ Zwei Teams von Datentypisten geben unabhängig voneinander jeweils zwanzig Spalten umfassende Wörterbuchabschnitte parallel ein (*double*

35 Vgl. zu den folgenden Ausführungen auch die Darstellung auf der Homepage des Projekts <http://gaer27.uni-trier.de/GrimmWB/grimmwb.htm>.

36 Unter <http://gaer27.uni-trier.de/MWV-online/MWV-online.html> finden sich weitere Informationen zu dem Projekt „Mittelhochdeutsche Wörterbücher auf CD-ROM und im Internet“. Für die Datenerfassung in China sprachen mehrere Gründe: Zum einen ist die Erfassung von Texten durch Nichtmuttersprachler weit weniger fehleranfällig, da der Faktor des „verstehenden“ und daher unbewusst korrigierenden Lesens wegfällt. Zum anderen ist mit Jingning Tao von der Trierer Arbeitsstelle des Projekts „Neues Mittelhochdeutsches Wörterbuch“, der den Kontakt zu der Eingabefirma vermittelt hat, ein direkter Ansprechpartner vor Ort.

keying), damit beim automatischen Vergleich der beiden Fassungen schon der größte Teil der Eingabefehler gefunden und anschließend korrigiert werden kann.

Die Eingabe erfolgt in WORD, weil so eine (Sofort-)Kontrolle der typographischen Merkmale am Bildschirm stattfinden kann. Die Eingabezeilen entsprechen jeweils den Zeilen im Wörterbuch. Verschiedene Kodierungen, die die spätere Auszeichnung vereinfachen, werden bereits jetzt angebracht; z. B. werden Anfang und Ende von Verszitäten mit ## und ##{ markiert, Lemmata werden durch #F+ und #F-, der Kodierung für fett, ausgezeichnet, Einschübe bei Verszeilen haben eindeutige Kodierungen wie &t0 &t0 &DUAN& für Zeilenüberläufe, Belegstellenangaben in Verszitäten werden in #CC+ und #CC- eingeschlossen, wie die folgende Gegenüberstellung zeigt:

GNISCHT, n., *dasselbe wie gnisch und ⁴gnist, s. unter genist 2 a und 3 a teil 4, 1, 2, 3472f., auch FISCHER schwäb. 3, 363.*

GNISKES, gnisk, m., 'geizhals': er (*ihr*) saht, doss ich kee knauser oder gniskes bin GRYPHIUS *lustsp.* 311 *Palm*;
Gniscus thut niemanden nichts; dennoch ist ihm niemand

gut,

eben darum, weil er nie keinem etwas gutes thut

LOGAU *sinnged.* 543 *Eitner*;

vgl. gnisk, als schimpfwort, in einer sprichwortsammlung aus dem anfang des 17. jh. bei FRISCHBIER preusz. 1, 242^b.

649.34 \$0 #F+gnischt,#F- #/+n., dasselbe wie#/- gnisch #/+und#/- #H:4#G:gnist,
#/+s. unter#/-
649.35 \$0 genist 2 a #/+und#/- 3 a #/+teil#/- 4, 1, 2, 3472#/+f., auch#/-
#k+Fischer#k-
649.36 \$0 #/+schwäb.#/- 3, 363.
649.37 \$0 #F+gniskes,#F- gnisk, #/+m.,#/- '#/+geizhals#/-' #/+:#/- er
(#/+ihr#/-) saht, doss ich
649.38 \$0 kee knauser oder gniskes bin #k+Gryphius#k- #/+lustsp.#/- 311
#/+Palm;#/-
649.39 \$0 ##Gniscus thut niemanden nichts; dennoch ist ihm niemand
649.40 \$0 &t0 &t0 &DUAN&gut,
649.41 \$0 eben darum, weil er nie keinem etwas gutes thut
649.42 \$0 @/ #CC+#k+Logau#k- #/+sinnged.#/- 543 #/+Eitner;#/-#CC-##{
649.43 \$0 #/+vgl.#/- gnisk, #/+als schimpfwort, in einer
sprichwortsammlung
649.44 \$0 aus dem anfang des#/- 17. #/+jh. bei#/- #k+Frischbier#k-
#/+preusz.#/- 1, 242#H:b#G:.

Abb. 2: Teil der Strecke GLIBBER-GRÄZIST (Bd. 8) im Original und im TUSTEP-Format³⁷

37 Die Bandzählung bezieht sich auf die Taschenbuchausgabe des DWB [Nachdruck] von 1984.

Im Februar 1999 wurde mit der Eingabe des DWB begonnen; zur Zeit (Januar 2000) liegen 18 Bände, d.h. die Bände 1 (A – Biermolke), 2 (Biermörder – Dwatsch), 3 (E – Forsche), 4 (Forschel – Gefolgsman), 5 (Gefoppe – Getreibs), 6 (Getreide – Gewöhnlich), 7 (Gewöhnlich – Gleve), 8 (Glibber – Gräzist), 19 (Stob – Stollen), 22 (Treib – Tz), 25 (V – Verzwunzen), 26 (Vesche – Vulkanisch), 28 (W – Wendunmut), 29 (Wenig – Wiking), 30 (Wilb – Ysop), 31 (Z – Zmasche), 32 (Zobel – Zypressenzweig), 33 (Quellenverzeichnis) komplett maschinenlesbar vor. Die mit WORD erfassten Daten werden per E-Mail von China nach Trier geschickt und über WordPerfect nach TUSTEP konvertiert. Dabei werden auch die Datensätze der jeweiligen Spalte und Zeile entsprechend nummeriert, damit der erfasste Text eine Referenz zum gedruckten Wörterbuch enthält (vgl. Abb. 2). Beide Versionen einer zwanzigspaltigen DWB-Strecke werden danach mit VERGLEICHE automatisch in TUSTEP verglichen; auf der Grundlage des daraus entstehenden Vergleichsprotokolls, aus dem die Abweichungen zwischen den beiden Dateien abgelesen werden können, werden die Daten nunmehr von Hand korrigiert. Das Ergebnis nach der Zusammenführung der einzelnen Dateien durch das TUSTEP-Kommando LISTE ist eine weitgehend fehlerfreie Version der jeweiligen Bände bzw. Wortstrecken des DWB, die zusätzlich durch verschiedene Fehlersuch-Routinen bereinigt wird.³⁸ Schon die korrigierte TUSTEP-Datei kann mit einfachen Zeigefunktionen zur Datenanalyse herangezogen werden, auch ist es schon jetzt möglich, mit KOPIERE Listen mit Informationen aus dem Datenbestand zu erstellen.³⁹

Für die strukturelle Auszeichnung des DWB wird die streng hierarchisch organisierte Auszeichnungssprache SGML (*Standard Generalized Markup Language*) verwendet, da die metasprachliche Kodierung der Artikelteile eine plattformunabhängige Verarbeitung der Daten garantiert. Diese sind sowohl für die Herstellung der Internet-Version als auch für die Umsetzung in eine CD-ROM verwendbar. Anstelle selbstdefinierter Tags wurden von Anfang an die besonders für philologische Zwecke von der TEI (*Text Encoding Initiative*) vorgeschlagenen Kodierungen und die speziell für lexikographische Werke von der TEI vorgesehene DTD (*Document Type Definition*) genutzt.⁴⁰ Weitere SGML-konforme Regeln für die Kodierung von Sonderzeichen (*SGML entity sets*), die für die Digitalisierung des DWB in großer Zahl benötigt werden, ergänzen die Bearbeiter jeweils in Absprache mit demselben Informatiker, der auch das Projekt „Mittelhochdeutsche Wörterbücher auf CD-ROM und im Internet“ betreut.

Das KOPIERE ist, in Verbindung mit den üblichen Pattern-Matching-Funktionen, der wichtigste Baustein für die Auszeichnung. Dabei ist die Arbeit mit Wahlschaltern von besonderem Interesse, da es aufgrund der Zeichenbegrenzung in TU-

38 Vgl. zu Fehlersuche und Sonderzeichenbehandlung ausführlich den 3. Teil dieses Beitrags.

39 Vgl. hierzu auch den Beitrag von JOHANNES FOURNIER in diesem Band.

40 Zur TEI vgl. unter anderen IDE/SPERBERG-MCQUEEN (1995) und <http://www.uic.edu/orgs/tei>.

STEP beim Parameter AA, der das Zusammenfassen von Datensätzen steuert, ansonsten nicht möglich wäre, zeilen- bzw. datensatzübergreifend nach bestimmten Kodierungen zu suchen.⁴¹

Da einzelne Artikelteile nicht allein aufgrund der bei der Eingabe bereits kodierten strukturellen Merkmale automatisch gegeneinander abgegrenzt werden können, orientiert sich die Auszeichnung insbesondere an typographischen Merkmalen. Autorangaben z. B. können problemlos mit öffnenden und schließenden Tags versehen werden, da sie immer in Kapitälchen stehen. Allerdings ist die Typographie innerhalb des DWB sehr uneinheitlich, was oft erst nach einer probeweisen Auszeichnung deutlich wird. So stellt sich vor allem die mehrdeutige Funktion der kursiv im Wechsel mit recte gesetzten Passagen noch als Problem dar, sowohl für die Analyse als auch für die Auszeichnung des DWB, wie das folgende Beispiel zeigt: Belegstellenangaben mit Werktitel, aber ohne Autorangabe sind lediglich durch Kursive (die normale typographische Darstellung von Werktiteln im DWB) eingeleitet. Kursives kann jedoch außer bei in Belegstellenangaben vorkommenden Werktiteln an nahezu jeder Stelle innerhalb eines Artikels stehen; daher ist die Auszeichnung von solchen Belegstellenangaben, die nicht durch Autornamen, sondern Werktitel eingeleitet sind, besonders schwierig.

Um diese Schwierigkeiten frühzeitig in den Griff zu bekommen, wurde das Quellenverzeichnis (Band 33) vorrangig maschinenlesbar gemacht und bearbeitet. Es kann daher jetzt an der Auszeichnung von Band 33, dem Quellenverzeichnis, und Band 8, der die Wortstrecke GLIBBER-GRÄZIST umfasst, parallel gearbeitet werden.⁴² Auf diese Weise können solche Belegstellen, die aus den oben genannten Gründen schwer zu erfassen sind, durch die Verlinkung mit dem Quellenverzeichnis gefunden bzw. ausgezeichnet werden.

Das manuelle Eingreifen beim Einfügen der SGML-Kodierungen soll möglichst vermieden werden, um zu gewährleisten, dass die Auszeichnung immer auf denselben Daten aufbaut. Daher werden die Dateien bislang allein mit Hilfe der jeweils erarbeiteten TUSTEP-Routinen und ohne manuelle Eingriffe ausgezeichnet. Durch die uneinheitliche Artikelstruktur und Typographie im DWB ist es immer wieder erforderlich, die Auszeichnungsroutinen zu modifizieren. Die Dokumentanalyse geht so Hand in Hand mit der Erarbeitung der Auszeichnungsroutinen.

Auf der höchsten Hierarchie-Ebene lassen sich mittlerweile alle Artikel zuverlässig auszeichnen, bei Belegstellenangaben und Verszitaten auch auf tieferen Ebenen. So können bisher sukzessive mit Tags versehen werden: Alle Einträge,⁴³ Lemmata, die jeweils mit einer eindeutigen Identifikationsnummer versehen werden,

41 Der längste Artikel des DWB (Lemma GOTT) umfasst 127 Spalten; das ergibt bei 4500 Zeichen pro Spalte 5711 500 Zeichen und nach Einsetzen der SGML-Kodierungen natürlich eine noch wesentlich höhere Zahl.

42 Eine erste Umsetzung des achten Bandes wird im Laufe des Jahres 2000 auf der Homepage des Projekts vorgestellt werden.

43 Die Reihenartikel werden auch als solche ausgezeichnet.

Lemmavarianten, Sublemmata, grammatische Angaben, Gliederungsmarken und Kapitelgliederungstext mit dem Abschnitt, auf den sich die Gliederungsmarke bezieht, und Verszitate mit Belegstellenangaben. Ausgehend von der obersten Ebene werden mit zunehmender Auszeichnung auch Elemente niedrigerer Hierarchieebenen mit Tags versehen, wie die folgende Abbildung verdeutlicht:

```

<entry id='GG02393' n='649.34'>
  <form type='lemma'>gnischt</form>
  <gramgrp>
    <gram type='n'>n.</gram></gramgrp>
    <sense>#/+, dasselbe wie#/- gnisch #/+und##H:4#G:gnist, #/+s. unter#/- genist 2 a
    #/+und#/- 3 a #/+teil#/- 4, 1, 2, 3472#/+f., auch#/-</sense>
    <sense>
      <title type='sigle'>
        <bibl>
          <author>Fischer</author> #/+schwäb.#/-</bibl>
          <ref>3, 363.</ref></title></sense>
        </entry>
      <entry id='GG02394' n='649.37'>
        <form type='lemma'>gniskes</form>
        <form type='variant' rend=', '>gnisk</form>
        <gramgrp>
          <gram type='m'>m.</gram></gramgrp>
          <sense>#/+,#/- '#/+geizhals#/-': er (#/+ihr#/-) saht, doss ich kee knauser oder gniskes
          bin</sense>
          <sense>
            <title type='sigle'>
              <bibl>
                <author>Gryphius</author> #/+lustsp.#/-</bibl>
                <ref>311</ref></title></sense>
                <sense>#/+,#/- '#/+Palm;#/-</sense>
                <sense><add rend='vers'>
                  <q type='line'>Gniscus thut niemanden nichts; dennoch ist ihm niemand gut, </q>
                  <q type='line'>eben darum, weil er nie keinem etwas gutes thut</q></add></sense>
                <sense>
                  <title type='sigle'>
                    <bibl>
                      <author>Logau</author> #/+sinnged.#/-</bibl>
                      <ref>543</ref></title></sense>
                      <sense>#/+,#/- '#/+Eitner;#/-</sense>
                      <sense>#/+,#/- '#/+vgl.#/- gnisk, #/+als schimpfwort, in einer sprichwortsammlung aus dem an-
                      fang des#/- 17. #/+jh. bei#/-</sense>
                    <sense>
                      <title type='sigle'>
                        <bibl>
                          <author>Frischbier</author> #/+preusz.#/-</bibl>
                          <ref>1, 242#H:b#G:.</ref></title></sense>
                        </entry>

```

Abb. 3: Teil der Strecke GLIBBER-GRÄZIST (Bd. 8) im TUSTEP-Format mit SGML/TEI-konformen Kodierungen.

Nach der Auszeichnung der Dateien mit SGML/TEI-konformen Kodierungen werden diese von den Mitarbeitern unter Benutzung eines Standard-SGML-Parsers validiert.

Die Routinen für die Umsetzung in die CD-ROM- und die Internetversion werden durch einen Informatiker erstellt. Eine Aufbereitung der kodierten Daten für die Internet- bzw. CD-ROM-Version ist jederzeit möglich und kann somit als Kontrolle für die Auszeichnungsarbeit unmittelbar herangezogen werden. Das elektronische Wörterbuch wird in übersichtlicher Form auf dem Bildschirm angezeigt, zugleich aber soll über PostScript- bzw. PDF-Files am Bildschirm auf eine Version zugegriffen werden können, die das Wörterbuch in seiner gedruckten Fassung simuliert. Diese Druckfassung wird aus denselben Ausgangsdaten wie die Internet- bzw. CD-ROM-Version generiert; dabei werden die PostScript-Files mit Hilfe des TUSTEP-Satzprogramms erstellt.

3. Zur Fehlersuche und Sonderzeichenbehandlung in der Datengrundlage

Ein Arbeitsschritt, der die Herstellung einer qualitativ hochwertigen Datengrundlage für das elektronische Wörterbuch wesentlich betrifft, in der Regel aber wenig Beachtung in der Dokumentation findet, ist die Fehlersuche. Um fehlerfreie⁴⁴ Dateien zu erzeugen, wird daher nach Abarbeitung der Vergleichsprotokolle eine Reihe weiterer Arbeitsschritte notwendig, die sich verschiedenen Aspekten der Fehlersuche und Problembehandlung im Datenbestand widmen und die im Folgenden näher beschrieben werden sollen.

Einfache Eingabefehler werden, wie bereits oben in Abschnitt 2 dargestellt, durch den Abgleich der doppelten Eingabe behoben (Vergleichsprotokoll des automatischen Vergleichs von Eingabeversion A und B). Neben solchen einfachen Tipp- und Eingabefehlern wird über das Vergleichsprotokoll eine Reihe weiterer typographischer Problemfälle erfasst, die bereits bei der Eingabe in China als solche erkannt und von den Datentypisten markiert werden. Als erster Schritt erfolgt die qualifizierte Weiterbearbeitung dieser markierten typographischen Besonderheiten und Sonderzeichen.

Allein schon aufgrund seiner Anlage als Belegwörterbuch findet sich im DWB eine Vielzahl von nicht alltäglichen Schriftzeichen und graphischen Symbolen. Eine große Gruppe davon bilden die Zitate und Belege aus nicht-lateinischen Alphabeten. Besonders häufig finden sich griechische und hebräische Belege.

⁴⁴ Fehlerfreiheit kann nur als Ziel gesetzt, aber in einem Datenbestand dieser Größe nicht empirisch nachgewiesen werden. Da Fehlerfreiheit in einem solchen Datenbestand also nicht verifizierbar ist, kann eine Datei rein technisch nicht als „fehlerfrei“ bezeichnet werden. Der Einfachheit halber wird diese Bezeichnung hier beibehalten, auf ihre Problematik soll aber hiermit hingewiesen sein.

Griechische Textteile werden bei der Eingabe vollständig erfasst und müssen lediglich auf ihre Richtigkeit hin überprüft werden. Hierbei ist das Ergebnis des Vergleichsprotokolls in der Regel ausreichender Anhaltspunkt. Anders bei den hebräischen Textteilen, die bei der Eingabe nicht erfasst werden und daher nachgetragen werden müssen; dies setzt ein Potenzial an sprachlichen Kompetenzen voraus, die außerhalb des Projekts in Zusammenarbeit mit dem Fach Jiddistik an der Universität Trier zur Verfügung gestellt werden.

Neben Griechisch und Hebräisch finden sich Belege aus weiteren, teilweise nicht mehr allgemein bekannten Sprachen wie Altkirchenslavisch, Avestisch, Estnisch, Finnisch, Lettisch, Lappisch usw. In den etymologischen Anschlüssen gehen die Wörterbuchbearbeiter in der Regel bis auf die indogermanische Stufe des Worts zurück, was zur Folge haben kann, dass Belege aus den entlegensten Ästen des indogermanischen Sprachenstammbaums angeführt werden. Uneinheitliche Notationssysteme sowie Fremdalphabete bedingen an solchen Stellen des DWB typographische Problemfälle. So ist beispielsweise in Belegen aus dem Gotischen die Ligatur ⟨h + v⟩ vertreten, oder für das Altnordische das spirantisierte *b* mit einem Querstrich durch den Schaft. Beide Sonderzeichen sind mit den vorhandenen TU-STEP-Zeichensätzen (noch) nicht bzw. nicht korrekt⁴⁵ darstellbar. In diesen Fällen werden Stellvertreterzeichenfolgen definiert, die an die Stelle der Sonderzeichen in die Datei gesetzt werden und zunächst eine reine Platzhalterfunktion haben. Diese Stellvertreter haben eine beliebige, aber immer gleiche Grundform, damit sie systematisch auffindbar sind. Für die DWB-Daten werden als Ersatzkodierungen immer ein oder mehrere Zeichen in geschweiften Klammern verwendet. Diese immer gleiche Grundform ermöglicht eine Weiterbearbeitung dieser Sonderzeichen zu einem beliebigen späteren Zeitpunkt. Es ist rationeller, die vollständige Liste nicht darstellbarer Zeichen nach Ende der Eingabephase geschlossen weiter zu verarbeiten, als jedes anfallende nicht darstellbare Zeichen gesondert zu behandeln.

In den etymologischen Artikelteilen des DWB, die in jüngeren Wörterbuchlieferungen häufig unter der Rubrik *herkunft und form* abgesetzt sind, findet sich generell eine Vielzahl von weiteren Zeichen, die vor Abbildungsschwierigkeiten stellen. Eine besondere größere Gruppe hierbei sind die Belege aus den großen deutschen Dialektwörterbüchern. Deren Spezialität sind uneinheitliche, oft idiosynkratische Notationssysteme; zur Zeit der Abfassung der meisten vorliegenden Mundartwörterbücher des deutschen Sprachraums gab es keine vereinheitlichenden Standards wie z. B. die heute der wissenschaftlichen Dialektologie zur Verfügung stehenden phonetischen Transkriptionsalphabete der *International Phonetic Association*.⁴⁶ So hat im Grunde jeder Lexikograph für sein Wörterbuch ein eigenes No-

45 Das oben im Text zu sehende Reibelaut-*b* wurde mittels Hinzusetzen einer „Mittellinie“ (vgl. TU-STEP-Referenz 1993, S. 324) erzeugt, ist aber als Ersatzdarstellung inakzeptabel, da der Strich den Schaft durchschneiden soll und nicht den Buchstabenkörper.

46 Vgl. <http://www2.arts.gla.ac.uk/IPA/ipa.html>.

tationssystem entworfen. Bei diesen Belegen aus Dialektwörterbüchern stellen insbesondere die Diakritika ein Problem dar: Kombinationen von Diakritika sowie heute ungebräuchliche Diakritika strapazieren die in Zeichensätzen vorhandenen Möglichkeiten. Das DWB bewahrt die Notation, soweit nachprüfbar, in aller Regel akribisch; bei der Übernahme in den Datenbestand kann nur versucht werden, diese unter TUSTEP zunächst so genau wie möglich nachzubilden, wenn notwendig wiederum durch Stellvertreterzeichenfolgen. Die Abbildung dieser Sonderzeichen unter SGML/HTML-Bedingungen stellt einen Arbeitsschritt dar, der zu einer späteren Bearbeitungsphase gehört und daher an dieser Stelle nicht weiter erörtert werden soll.

Eine weitere spezifische Sonderzeichengruppe sind ältere, heute nicht mehr gebräuchliche typographische Zeichen und Symbole z. B. für Währungen und Gewichte wie etwa die Zeichen für *Pfund*, *Mark*, *Schilling* oder *Pfennig*. Zu dieser Gruppe gehören auch weitere typographische Merkwürdigkeiten, die dann und wann auftauchen können, mitunter sogar nur ein einziges Mal im DWB vorkommen, so beispielsweise eine Zickzacklinie in einem Beleg, der sich unter dem Lemma „Zickzack“ (Bd. 31, Sp 887) findet. Zum einen muss oftmals zunächst die Bedeutung eines solchen Zeichens geklärt werden, zum anderen fehlen diese Zeichen größtenteils in vorhandenen Zeichensätzen; hier werden wiederum Stellvertreterzeichen notwendig.

Ein zweiter Arbeitsschritt ist die Behandlung von Druckfehlern. Da auch das DWB nicht frei von Druckfehlern ist, wird es notwendig, diese, soweit sie aufgefunden werden können, nicht nur zu korrigieren, sondern auch zu markieren. Eine systematische Suche nach Druckfehlern gestaltet sich allerdings schwierig, nur ein bestimmter Prozentsatz kann durch automatisierte Suchroutinen aufgefunden werden. Dies ist etwa der Fall bei unwahrscheinlichen Buchstabenkombinationen innerhalb eines Worts. So kann beispielsweise systematisch nach drei gleichen Buchstaben, die hintereinander stehen, mittels der Pattern Matching-Funktionen von TUSTEP gesucht werden. Nach Anwendung eines im Projekt für das DWB entwickelten Fehlersuchprogramms konnten z. B. folgende Druckfehler ermittelt werden: *götterhallee*, *himmel*, *gottesschau* (alle aus Bd. 8). Bei der Kontrolle in der gedruckten Vorlage (um auszuschließen, dass es sich um Eingabefehler handelt) finden sich diese Dreierkonsonanzen schon so. Zu beachten ist bei solchen Dreierkonsonanzen aber, dass sie in Komposita wie *Schiffahrt* durchaus als reguläre Schreibungen vorkommen können. Aufgefundene tatsächliche Druckfehler jedoch werden korrigiert und zusätzlich mit der Markierung {!D} versehen.

Andere Druckfehler können, wenn sie nicht korrigiert sind, womöglich die Ergebnisse der Weiterverarbeitung verfälschen. So findet sich z. B. in einer Sigle der Autorennamen „SEGHERS“. Das kursive *s* kann als Genitiv-*s* missverstanden werden. Eklatanter kann sich aber ein solcher Druckfehler, wenn er nicht rechtzeitig aufgefunden wird, bei der späteren automatischen Auszeichnung der Siglen auswirken,

denn hierbei dienen die Auszeichnungen für Kapitalchen als Begrenzungsmarken für die Autorennamen. Durch die automatische Austauschprozedur wäre folglich der Name SEGHERS zu *SEGHER verkürzt.

Nicht jeder gefundene Druckfehler ist zweifelsfrei zu berichtigen. Einzelne beim Druck ausgefallene Zeichen, beispielsweise einzelne Zahlen bei Stellenangaben, sind nicht immer intuitiv oder durch Nachweis (z. B. durch Konsultation des Quellenverzeichnisses zum DWB) rekonstruierbar. Ebenso ist nicht jede Abweichung von der Normalorthographie eindeutig als Druckfehler klassifizierbar.

Solche vermeintlichen und/oder nicht auflösbaren Druckfehler werden gesondert behandelt und in der Datei durch {?D} gekennzeichnet. Sämtliche aufgefundenen Druckfehler werden zusätzlich in eine fortlaufende Liste aufgenommen, womit ohne zusätzlichen Aufwand eine Errata-Liste für das DWB entsteht.

Tipp- und Eingabefehler werden durch die Methode der doppelten Eingabe und des automatischen Abgleichs über das Vergleichsprotokoll verbessert. Selten allerdings kommt es zu Eingabefehlern, die simultan in beiden Versionen auftreten und daher nicht als Abweichung einer Version von der anderen erkannt werden können. Dieses sporadische Phänomen ist allerdings beschränkt auf sehr eingegrenzte Umgebungsbedingungen: So wird z. B. ein *i*, welches in Ligatur mit vorangehendem *f* steht, als *l* gelesen, oder umgekehrt das *l* als *i*, dies aber nur, wenn als weitere Umgebungsbedingung die kleinste im DWB verwendete Fontgröße hinzukommt, die eine kognitive Zeichenerkennung zusätzlich erschwert. So finden sich in Band 8 folgende in beiden Eingabeversionen aufgetretene Eingabefehler: *flnstern* statt *finstern* und ähnlich: *fiugs*, *fleng*, *hofleren*, *gefleder*, *flnger*, *flsch*, *Hafls*. Eine ähnliche Fehlerquelle ist die Verwechslung von *e* und *c*: *wohlbckannten*, *entbchren* (alle aus Bd. 8). Die *e-c*-Verwechslung wird mit folgendem Suchlauf abgeprüft:

```

>1z      aeiouäöüy#%
>2z      ., ; : ! ?)
ZF+      |f|c||f|ch|f|cz||g|c||g|chen||j|c||j|ch||k|c||k|ch||
ZF+      |l|c||l|ch|l|ck||
ZF+      |m|c||m|ch|m|c||
ZF+      ||n|ch|n|c>1|n|ck>1|n|c1.|>1|n|c|n|cz|n|c>2|n|c||n|c|
ZF+      |p|c||p|ch|p|ck|q|c|
ZF+      ||s|ch|s|cr|s|cl|s|c>1|s|c.|s|c||s|c|
ZF+      ||t|ch|t|ck|t|c.|t|c>1||t|c|v|c|w|c|
ZF+      ||e|x|c||x|c|
ZF+      ||z|ch|z|cw|z|c>1||z|c|
ZF+      |c|d|c|j|c|p|c|s|c|v|c|w|
ZF+      ||e|rc|g|a|rc|g|c|g|m.|c|g|l.|c|g|
ZF+      ||r|c|ma|c|m|c|me||c|m|
ZF+      ||c|ni||c|n|

```

Abb. 4: Ausschnitt aus einer Fehlersuchroutine.

In gleicher Weise können, bedingt durch die beim Satz des DWB verwendeten Griechisch-Fonts, die griechischen Buchstaben ϕ und ψ verwechselt werden. Solche Fehler sind durch automatisierte Prozeduren (TUSTEP-KOPIERE) sehr effektiv aufzufinden, da die eingeschränkten Umgebungsbedingungen des Vorkommens der Suche mit Hilfe von Pattern Matching entgegen kommen. Beachtet werden müssen beim Programmieren von Suchprozeduren für Dateien des elektronischen DWB allerdings die großen orthographischen Spielräume, welche bei einem Werk dieses Umfangs und einer sich über 100 Jahre erstreckenden Ausarbeitungszeit erwartet werden müssen, sich aber auch insbesondere durch die zeitliche und räumliche Belegvielfalt erklären.

Nach der Bearbeitung von Sonderzeichen, Druckfehlern und doppelten Eingabefehlern wird die Datengrundlage weiteren Konsistenzprüfungen unterzogen. Durch automatisierte Fehlersuche über TUSTEP-KOPIERE-Programmläufe lassen sich nicht nur die oben beschriebenen Druckfehler des DWB finden. Solche Suchprozeduren werden ebenso dazu genutzt, im Datenbestand vorhandene Auszeichnungen auf ihre Konsistenz hin zu überprüfen. So kann z. B. getestet werden, ob auf eine öffnende Markierung eine schließende Markierung folgt, oder ob sachliche Auszeichnungen, z. B. Kennzeichnungen von Verszitäten, die teilweise bereits während der Eingabe erfolgen, in richtiger Art und Weise angebracht sind. Diese in den TUSTEP-Dateien schon während der Eingabe angebrachten expliziten Auszeichnungen von Textteilen sind hilfreich und konstituierend für die in den späteren Arbeitsschritten erfolgende Kodierung der Daten nach SGML/TEI-Richtlinien; daher muss auf ihre Konsistenz ebenso ein prüfendes Augenmerk gerichtet werden.

Die erforderlichen Schritte zur Herstellung einer wirklich fehlerfreien Datengrundlage erschöpfen sich also nicht in einem einfachen automatischen Abgleich zweier Eingabeversionen. Auch der kleine danach verbleibende Fehlerrest macht die Etablierung von weiteren oben beschriebenen Schritten zur Fehlersuche und Problembehandlung notwendig und damit die Fehlersuche zu einem Prozess, der zu keinem Zeitpunkt der Datenerfassung als wirklich abgeschlossen angesehen werden kann.

Die Brauchbarkeit der elektronischen Version eines Wörterbuchs hängt ganz wesentlich ab von der Qualität der Datengrundlage. Daher sollte jede Möglichkeit – insbesondere jede Möglichkeit des automatisierten Fehlerbereinigungs – zur Optimierung der Datengrundlage wahrgenommen werden.

TUSTEP bietet für die oben beschriebenen Arbeitsschritte eine Reihe von Nutzungsmöglichkeiten, die von der Arbeit mit Pattern-Matching-Suchoperationen im Editor bis zum Einsatz des bewährten KOPIERE-Programm-Moduls zur qualifizierten und optimierten Fehlersuche reichen. Jedem der oben beschriebenen Erfordernisse zum Erstellen einer möglichst fehlerfreien Datengrundlage wird somit TUSTEP in ideal zu nennender Weise gerecht. Eine Reihe der in diesem Beitrag dargestellten Schwierigkeiten und bedenkenswerten Details begleitet also in allen Pha-

sen die Erarbeitung des digitalen DWB, die deshalb mitunter als eine schwierige Aufgabe erscheint; doch warum es nicht wie Jacob Grimm halten, der schon zu Beginn der Ausarbeitung des DWB sagte: „zuweilen möchte ich mich erheben und alles wieder abschütteln, aber die rechte besinnung bleibt dann nicht aus. Es gälte doch für thorheit, geringeren preisen obschon sehnsüchtig nachzuhängen und den groszen ertrag auszer acht zu lassen.“⁴⁷

Literatur

- BAHR, JOACHIM: Eine Jahrhundertleistung historischer Lexikographie: Das Deutsche Wörterbuch, begr. von J. und W. Grimm. In: Sprachgeschichte. Ein Handbuch zur Geschichte der deutschen Sprache und ihrer Erforschung. Hg. v. Werner Besch, Oskar Reichmann, Stefan Sonderegger. Erster Halbband. Berlin, New York 1984a, S. 492–501.
- BAHR, JOACHIM: Das Deutsche Wörterbuch von Jacob Grimm und Wilhelm Grimm. Stationen seiner inneren Geschichte. In: Sprachwissenschaft 9 (1984b), S. 387–455.
- BAHR, JOACHIM: Periodik der Wörterbuchbearbeitung. Veränderung von Wörterbuchkonzeption und -praxis. In: Kirkness/Kühn/Wiegand (Hg.): Studien zum Deutschen Wörterbuch, Bd. I, S. 1–50.
- BERGMANN, ROLF: *Projekt und Projektmacher*. Ein Beispiel für lexikographische Benutzerinteressen und lexikographische Befunde. In: Sprachwissenschaft 24 (1999), S. 337–360.
- BRAUN, WILHELM: Das Deutsche Wörterbuch seit seiner Übernahme durch die Akademie der Wissenschaften zu Berlin 1908 bis zu seinem Abschluss 1960. In: Dückert (Hg.): Das Grimmsche Wörterbuch. S. 125–152.
- Deutsches Wörterbuch von Jacob Grimm und Wilhelm Grimm. 16 Bde. [in 32 Teilbänden]. Leipzig 1854–1960. — Quellenverzeichnis 1971.
- DÜCKERT, JOACHIM (Hg.): Das Grimmsche Wörterbuch. Untersuchungen zur lexikographischen Methodologie. Stuttgart 1987a.
- DÜCKERT, JOACHIM: Jacob und Wilhelm Grimm. In: Ders. (Hg.): Das Grimmsche Wörterbuch, S. 7–48.
- DÜCKERT, JOACHIM: Das Deutsche Wörterbuch von Jacob Grimm und Wilhelm Grimm und seine Neubearbeitung. In: Jahrbuch der Henning-Kaufmann-Stiftung zur Pflege der Reinheit der deutschen Sprache 1986. Marburg 1987, S. 25–44.

47 JACOB GRIMM: Vorrede, S. III.

- HORLITZ, BERND: Deutsches Wörterbuch – Hausbuch der Nation? Probleme der Benutzung und Benutzungsmöglichkeiten. In: Kirkness/Kühn/Wiegand (Hg.): Studien zum Deutschen Wörterbuch, Bd. II, S. 407–434.
- HUBER, ANNA: Kritiker und Konkurrenten, erste Mitarbeiter und Fortsetzer der Brüder Grimm am Deutschen Wörterbuch. In: Dücker (Hg.): Das Grimmsche Wörterbuch, S. 49–90.
- IDE, NANCY/SPERBERG-MCQUEEN, C. M.: The TEI: History, Goals, and Future. In: Computers and the Humanities 29 (1995), S. 5–25.
- KIRKNESS, ALAN/KÜHN, PETER/WIEGAND, HERBERT ERNST (Hg.): Studien zum Deutschen Wörterbuch von Jacob Grimm und Wilhelm Grimm. 2 Bde. Tübingen 1991. [Lexicographica: Series maior; Bd. 33/34].
- KIRKNESS, ALAN/KÜHN, PETER/WIEGAND, HERBERT ERNST: Zur Einführung: Von der philologischen zur metalexikographischen Beschreibung und Beurteilung des Deutschen Wörterbuchs. In: Dies. (Hg.): Studien zum Deutschen Wörterbuch, Bd. I, S. VII–LXI.
- PÜSCHEL, ULRICH: Zwischen Erörterung und Ergebnisdarstellung. Zu Wörterbuchstilen im Deutschen Wörterbuch. In: Kirkness/Kühn/Wiegand (Hg.): Studien zum Deutschen Wörterbuch, Bd. I, S. 51–103.
- SCHLAEFER, MICHAEL: Zur Darstellung wortgeschichtlicher Zusammenhänge des 17.–20. Jahrhunderts in historischen Wörterbüchern. In: Sprachwissenschaft 24 (1999), S. 195–220.
- SCHRÖTER, ULRICH: Von Moriz Heyne zur Deutschen Kommission. Zur Bearbeitung des Deutschen Wörterbuchs von 1867 bis 1908. In: Dücker (Hg.): Das Grimmsche Wörterbuch, S. 91–124.
- SCHULZ, MATTHIAS: Der lexikographische Informationsgehalt in älteren Bedeutungswörterbüchern. Zugleich Überlegungen zum Nutzen einer Retrodigitalisierung älterer Wörterbücher. In: Sprachwissenschaft 24 (1999), S. 47–73.

