

Das Findebuch zum mittelhochdeutschen Wortschatz: Probleme der Kompilation von Ausgabenglossaren zu einem Gesamtglossar

Gärtner, Kurt; Hanrieder, Gerhard

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:
GESIS - Leibniz-Institut für Sozialwissenschaften

Empfohlene Zitierung / Suggested Citation:

Gärtner, K., & Hanrieder, G. (1993). Das Findebuch zum mittelhochdeutschen Wortschatz: Probleme der Kompilation von Ausgabenglossaren zu einem Gesamtglossar. *Historical Social Research*, 18(3), 125-130. <https://doi.org/10.12759/hsr.18.1993.3.125-130>

Nutzungsbedingungen:

Dieser Text wird unter einer CC BY Lizenz (Namensnennung) zur Verfügung gestellt. Nähere Auskünfte zu den CC-Lizenzen finden Sie hier:
<https://creativecommons.org/licenses/by/4.0/deed.de>

Terms of use:

This document is made available under a CC BY Licence (Attribution). For more information see:
<https://creativecommons.org/licenses/by/4.0>

die Vorbereitung der Sortierung werden nach *ux*. und etlichen anderen Abkürzungen vor dem nächsten großgeschriebenen Wort automatisch Nichtsortierzeichen eingefügt, weil in der Regel ein bei der Sortierung zu übergehender Vorname folgt. Wo diese Regel nicht greift, wird das Nichtsortierzeichen von Hand nachgetragen. Hierfür wird ein anderes Zeichen gewählt als das vom Programm für den gleichen Zweck vergebene, so daß man auch bei der Kontrolle schnell eine falsche Regel von einer falschen Einzelkorrektur unterscheiden kann.

4. In den Registern müssen neben Korrekturen auch Ergänzungen vorgenommen werden können. Um insbesondere automatische Ergänzungen durchführen zu können, muß man das fertige (vorläufige) Register wieder in einzelne Einträge zurückverwandeln und dann neu sortieren können.

Die hier eingeschlagene Strategie für die Zusammenarbeit zwischen einer mehr technisch orientierten Stelle einerseits und den Fachwissenschaftlern andererseits, die sich auf die sachlichen Fragen beschränken wollen und sollen, hat sich bei den Registern von Orts- und Personen-Namen bereits bestens bewährt. Prinzipiell die gleichen Regeln gelten natürlich auch für die Sachregister, auch wenn der Ausgangspunkt und die einzelnen Schritte der Arbeit sich in vielen Einzelheiten unterscheiden.

Das Findebuch zum mittelhochdeutschen Wortschatz. Probleme der Kompilation von Ausgabenglossaren zu einem Gesamtglossar.

*Kurt Gärtner (Trier); Gerhard Hanrieder (Erlangen)**

I.

Vor kurzem ist ein neuartiges lexikographisches Werk zum Mittelhochdeutschen, der von ca. 1050 bis ca. 1350 reichenden Periode der deutschen Sprache, erschienen, das sogen. >Trierer Findebuch<, das Ergebnis einer Gemeinschaftsarbeit der Trierer Altgermanisten:

Kurt Gärtner, Christoph Gerhardt, Jürgen Jaehrling, Ralf Plate, Walter **Roll**, Erika Timm. Datenverarbeitung: Gerhard Hanrieder: *Findebuch zum mittelhochdeutschen Wortschatz. Mit einem rückläufigen Index*. Stuttgart: Hirzel 1992.

Mit dem Findebuch soll die desolote Situation auf dem Gebiet der mittelhoch-

* Protokoll des 56. Kolloquiums über die Anwendung der EDV in den Geisteswissenschaften 21. November 1992

deutschen Lexikographie für zwei bis drei Jahrzehnte, genauer: bis zum Erscheinen eines neuen Sprachstadienwörterbuchs zum Mittelhochdeutschen, provisorisch überbrückt werden. Zwar sind in den letzten Jahren einige Autoren- und Werkwörterbücher sowie Spezialwörterbücher zu bestimmten Textsorten erschienen oder in Angriff genommen worden. Nicht nur Germanisten, sondern auch Historiker, Theologen und andere Forscher, die sich mit mittelhochdeutschen Texten beschäftigen, sind dennoch immer noch angewiesen auf zwei große, einander ergänzende Wörterbücher, die inzwischen weit über 100 Jahre alt sind:

1. das Mittelhochdeutsche Wörterbuch[<] von Benecke/Müller/Zarncke [= BMZ] (1854-1866, letzter Nachdruck 1990), und
2. das Mittelhochdeutsche Handwörterbuch[<] von Matthias Lexer [= Lexer] (1872-1878, letzter Nachdruck 1992).

Von diesen beiden Wörterbüchern ist nur der BMZ ein selbständiges Sprachstadienwörterbuch; der Lexer dagegen ein Handwörterbuch auf der Basis des BMZ, zu dem er auch als Supplement fungiert und schließlich noch als alphabetischer Index, denn die Lemmata des BMZ sind nicht streng alphabetisch, sondern nach Wortstämmen und ihren Wortfamilien angeordnet.

Die beiden alten Wörterbücher basieren vorwiegend auf Quellen aus der Zeit von 1150 bis 1230, insbesondere auf den wenigen in den drei Jahrzehnten um 1200 entstandenen >klassischen< Texten >Iwein<, >Parzival< und >Tristan<, die etwa ein Viertel der Belege stellen. Für das Mittelhochdeutsche in den zeitlichen Grenzen von 1050-1350 waren diese Wörterbücher von Anfang an nicht repräsentativ. Heute sind sie es noch weit weniger, denn die Quellenerschließung ist nach dem Abschluß des Lexer 1878 weitergegangen.

Das Findebuch hat daher die Aufgabe, die beiden Wörterbücher zu ergänzen, indem es die Lemmata der Ausgabenglossare zu den nach 1878 erschienenen Editionen mittelhochdeutscher Texte aus dem Zeitraum von ca. 1050 bis 1350/1400 kompiliert und damit die Informationen liefert, die ein gezieltes Nachschlagen und Auffinden der gesuchten Lemmata in den Einzelglossaren ermöglicht. Rund 90 der über 100 für das Findebuch herangezogenen Einzelglossare stammen aus Editionen von Texten, die vor und nach den drei Jahrzehnten um 1200 entstanden sind. Der Lemmabestand des Findebuchs (rund 35.000) erlaubt damit erstmals einen annähernd repräsentativen Überblick über den gesamten mittelhochdeutschen Wortschatz.

Das Findebuch ist kein selbständiges Wörterbuch, sein eingeschränkter Wert für den Benutzer kommt deshalb unmißverständlich in seinem Titel zum Ausdruck. Die Bezeichnung >Findebuch< war in Analogie zu Roy Wisbeys >finding lists</>Findelisten< gewählt worden, mit denen er das Auffinden aller Verbformen in den von ihm herausgebrachten Indizes und Konkordanzen ermöglichen wollte. Das Findebuch hat einen ganz ähnlichen Zweck; es dient lediglich als Wegweiser zu dem in den Ausgabenglossaren gebuchten Wortschatz und enthält nur soviel Informationen, wie zum raschen und sicheren Auffinden der

gesuchten Wörter nötig ist. Die Lemmata sind daher mit grammatischen Angaben und gegebenenfalls auch mit Bedeutungsangaben versehen, um Homographie zu disambiguieren; zusätzlich ist aber bei neuen, in den beiden alten Wörterbüchern noch nicht gebuchten Wörtern noch die Bedeutung angegeben, sofern diese nicht erkennbar ist aus den Teilen eines Wortes, das durch Komposition oder Derivation entstanden ist.

Für die Herstellung des Findebuchs war von vornherein die möglichst intensive Nutzung aller Möglichkeiten der EDV geplant, denn der für die Ausarbeitung vorgesehene Zeitraum war auf etwa fünf Jahre begrenzt. Alle mechanisierbaren niederen Arbeiten wie Sortieren, Alphabetisieren, Kontrollieren, Kompilieren u. ä. wurden zum größten Teil dem Computer überlassen, um die zeitlich aufwendige Präparierung der Glossarlemmata für die automatische Kompilation der Einzelglossare mit der größtmöglichen philologischen Sorgfalt durchführen zu können. Die einheitliche Präparierung oder Voredierung (preediting) der Glossarlemmata nach einer genauen Arbeitsanweisung war die Voraussetzung für die automatische Kompilation.

Die Ausgabenglossare waren nach sehr verschiedenen Prinzipien hergestellt. Es gab Glossare, die als Supplemente zu Lexers Handwörterbuch angelegt waren und daher ohne großen Aufwand präpariert werden konnten für die Eingabe ihrer Lemmata am PC. Es gab aber auch Glossare, die aus bloßen Wortformen mit Stellenangaben bestanden und eine aufwendige Voredierung verlangten, denn die Wortformen mußten im Text der Ausgabe nachgeschlagen werden, um den richtigen Lemmaformen zugeordnet werden zu können. Zahlreiche Glossare wiesen nicht normalisierte Lemmata auf, die vielfach nicht übereinstimmten mit den normalisierten Lemmaformen, nach denen die Artikel eines Wörterbuchs zu einer älteren Sprachstufe wegen der großen graphischen Varianz der Quellentexte angeordnet sein müssen. Es mußten z. B. Glossareinträge wie

grulich *adj.*, kapelan *stm.*

den für die Makrostruktur des Findebuchs maßgebenden Normalformen

griuwelich *adj.*, kappellän *stm.*

zugeordnet werden, d. h. die Glossarlemmata mußten lemmatisiert werden auf diese Normalformen hin. In den zitierten Beispielen bestand die Lemmatisierung nur in einer Korrektur des Glossareintrags, weil anhand der normalisierten Lemmaform des Findebuchs die etwas abweichende Form des Wortes im Ausgabenglossar ohne Schwierigkeiten aufzufinden war. Doch stark abweichende Glossarlemmata mußten im Findebuch als Sublemmata zu den Normalformen der Hauptlemmata angesetzt werden, um ihre Auffindbarkeit in den Ausgabenglossaren zu garantieren. Ebenso wie bei nicht übereinstimmenden Lemmaformen in Findebuch und Ausgabenglossar mußte auch bei differierenden grammatischen Angaben verfahren werden. Da im Findebuch die morphologische Information (über Flexionsart, Genus, Numerus usw.) der Ausgabenglossare möglichst genau registriert werden sollte, waren zusätzlich zu der

grammatischen Angabe beim Hauptlemma vielfach Unterangaben anzusetzen (z. B. häufig bei den Feminina, deren Flexion in der Geschichte des Deutschen einen erheblichen Wandel zeigt). So mußten z. B. zu dem Normalansatz, dem Findebuchhauptlemma, wie

triuwe *stf.*

nicht nur mehrere graphische Varianten des Lemmas angegeben werden, damit es in den Ausgabenglossaren auffindbar war, sondern auch die jüngere Flexionsart *swf.*, um die morphologische Information zu erhalten. Die Mikrostruktur des Findebuchartikels ist dadurch differenziert, wie das folgende Beispiel der Endfassung eines Artikels zeigt:

- triuwe *stf.* RWCHR. STRAMIS ...
- Triuwe **MINNER.I**
 - trüiwa **WILL.**
 - trüwē **L.ALEX. BRUN SCHÜREBR.**
 - Triuwe *swf.* **MINNER.I**
 - trüwec *swf.* **ETLH.**
 - *swf.* **PSM. SHORT**

Nur die philologisch anspruchsvolle Voredierung erlaubte es, aus der variantenreichen Masse der zu verarbeitenden Daten ein effizientes und verlässliches lexikographisches Hilfsmittel zu erstellen. Mit einer Beschränkung auf das aus dem Schriftbild der Glossarlemmata Ablesbare und dessen Alphabetisierung wäre kaum einem Benutzer gedient gewesen.

Durch die Kompilation der voredierten Einzelglossare (dazu s. u.) wurden immer wieder Zwischenstufen des stetig anwachsenden Findebuchs erzeugt, die gründlich revidiert wurden. Die Ergebnisse dieser Zwischenrevisionen waren in mehrfacher Hinsicht für das Projekt förderlich (s. u.); das nach einer Zwischenrevision ausgedruckte vorläufige Findebuch bot dem Geldgeber (Deutsche Forschungsgemeinschaft) eine zuverlässige Bilanz der geleisteten Projektarbeit und zugleich die Entscheidungsgrundlage für die Weiterförderung.

Das Findebuch hat durch die beschriebenen lexikographischen Bemühungen eine andere Qualität erreicht als die vielen bis weit in die 1970er Jahre produzierten und leider auch publizierten Indizes und Konkordanzen, die mit Hilfe der EDV hergestellt wurden. Dem Findebuch merkt man zunächst kaum an, daß EDV-Routinen alle Arbeitsschritte begleiteten und daß ohne das Werkzeug EDV vermutlich wesentlich mehr als fünfeinhalb Jahre für seine Ausarbeitung nötig gewesen wären. Es gab unter den an der Ausarbeitung Beteiligten anfangs wohl eine intensive Diskussion über Art und Umfang des Computereinsatzes für das Projekt, in der auch Befürchtungen und Vorbehalte geäußert wurden. Die Vielseitigkeit des Werkzeuges EDV war daher schon in der Anfangsphase unter Beweis zu stellen.

Die Abneigung gegen computererzeugte lexikographische Hilfsmittel ist nur zu verständlich, wenn man die bereits erwähnten Indizes und Konkordanzen

betrachtet, die - von rühmenswerten Ausnahmen abgesehen - das Ergebnis schneller Rechenanlagen und mangelnden lexikographischen Sachverstands waren. Heute hat sich die Ansicht allgemein durchgesetzt, daß elektronisch hergestellte Indizes und Konkordanzen ohne kompetente lexikographische Überarbeitung nur Hilfsmittel von beschränktem Wert sind und daher keineswegs für eine Veröffentlichung taugen. - Auf die verschiedenen EDV-Routinen, welche die Arbeiten am Findebuch von Anfang an begleiteten, soll nun im folgenden näher eingegangen werden.

II.

Im Anschluß an die Voredierung der Glossare erfolgte deren maschinelle Verarbeitung. Für jedes Ausgabenglossar ergaben sich folgende EDV-Arbeits-schritte:

- Datenerfassung
- Korrektur
- Automatisches Hineinkompilieren in die Findebuchdatei
- Archivierung.

Die Arbeiten wurden fast ausschließlich mit TUSTEP durchgeführt. Lediglich bei der Dateneingabe, die auf alle Fälle dezentral am PC erfolgen sollte, haben wir in den ersten Projektjahren vor der Verfügbarkeit von PC-TUSTEP auf andere Textverarbeitungsprogramme zurückgegriffen; die Weiterverarbeitung der damit erzeugten ASCII-Dateien erfolgte nach deren Umwandlung in TUSTEP-Format auf dem Zentralrechner des Trierer Rechenzentrums. Für die Korrekturarbeiten hat sich die Entscheidung als sehr fruchtbar erwiesen, jedes zu erfassende Glossar doppelt von zwei verschiedenen Mitarbeitern eingeben und mittels TUSTEP-VERGLEICHE automatisch ein Vergleichsprotokoll der beiden Versionen erzeugen zu lassen. Unter Auswertung dieses Vergleichsprotokolles wurde eine korrigierte Version der Eingabedatei erstellt. Ein formatierter Ausdruck dieser Datei wurde anschließend nochmals gegen die Voredierung Korrektur gelesen, um die wenigen Fälle zu beseitigen, in denen beim doppelten Eingeben identische Fehler gemacht wurden. Am Ende dieser Korrekturschritte stand somit jeweils eine Glossardatei, die sowohl philologisch als auch das Datenformat betreffend für die Aufnahme in das Findebuch präpariert war.

Das Kernstück der Datenverarbeitung stellte - dem Projektziel »Kompilation von Einzelglossaren zu einem Gesamtglossar« entsprechend - ein Programm zur automatischen Kompilation der Einzelglossardateien zum Gesamtglossar dar, genauer gesagt: ein Programm zum automatischen Hineinkompilieren jeweils einer Glossardatei in die Findebuchdatei. Das Programm, das dieses Hineinkompilieren leistet, hat für jeden Artikel der Glossardatei zu prüfen, ob der Artikel bereits in der Findebuchdatei vorhanden ist. Ist dies der Fall, so ist die durch die Glossardatei neu hinzukommende Information in den Findebuchartikel zu integrieren, andernfalls ist der noch nicht im Findebuch gebuchte Artikel

unverändert ins Findebuch zu übernehmen. Realisiert ist dies als TUSTEP-Programm, das zunächst Findebuchdatei und Glossardatei mittels Sortieren und Mischen in eine Datei zusammenführt, in der Artikel mit gleichlautendem Artikelnamen aufeinanderfolgend stehen. Das Zusammenführen gleichlautender Artikel leistet ein auf dieser Zwischendatei operierendes in TUSTEP-KOPIERE geschriebenes Programm, das die neue Glossarinformation in den Findebuchartikel einfügt, wobei zuerst Vorhanden- bzw. Nichtvorhandensein von Sublemmata und -angaben geprüft werden muß, um entscheiden zu können, an welcher Stelle des Findebuchartikels die neue Information eingefügt werden soll.

Mit diesem Programm konnte die Kompilation der Glossardateien voll automatisiert werden, und das durch die Kompilation der beiden ersten Glossare entstandene Findebuch konnte unmittelbar nach Abschluß der Vordierungs- und Korrekturarbeiten an einem Glossar um dieses erweitert werden. Diese inkrementelle Vorgehensweise bot frühzeitig die Möglichkeit, Inkonsequenzen zu erkennen, die sich bei der philologischen Bearbeitung durch verschiedene Mitarbeiter nicht ganz vermeiden ließen. Das Findebuch wurde deshalb, wie bereits erwähnt, mehreren Zwischenrevisionen unterzogen, die neben den dabei vorgenommenen Korrekturen auch den positiven Effekt hatten, daß sich in zunehmendem Maße Vorstellungen über Artikel struktur und Layout entwickelten und die dafür nötigen Programmentwicklungen - wie z. B. Sortierprogramme für die Artikelbinnenstruktur - frühzeitig parallel zu den restlichen Projektarbeiten in Angriff genommen werden konnten.

Im Zusammenhang mit einer Zwischenrevision entstand auch der Wunsch nach einem rückläufig sortierten Index zum Hauptglossar, der sich als überaus nützlich zusätzliches Hilfsmittel erwies und deshalb letztlich mit in die gedruckte Ausgabe aufgenommen wurde. Diese Entscheidung fiel um so leichter, als mit dem einmal erstellten Programm zur vollautomatischen Generierung des rückläufigen Index aus dem Hauptglossar keinerlei zusätzliche Arbeit anfiel, wenn aus neuen Findebuchversionen ein jeweils aktueller rückläufiger Index erzeugt werden sollte.

Bei der in der Endphase des Projektes in Bezug auf die Datenverarbeitung im Vordergrund stehenden Vorbereitung der Satzherstellung konnten die logischen Auszeichnungen der Dateien weitgehend automatisch in Steueranweisungen des Satzprogramms von TUSTEP ausgetauscht werden. Die hier skizzierte computerunterstützte Durchführung eines lexikographischen Projektes hat neben der von uns intensiv genutzten Möglichkeit, korrekturunterstützende automatische Konsistenzüberprüfungen durchführen zu lassen, den positiven Seiteneffekt, daß am Ende des Projektes neben der gedruckten eine maschinenlesbare Version des entstandenen Werkes zur Verfügung steht, die zusätzliche Recherchemöglichkeiten eröffnet.