Wörterbuch und Grammatik als Folgeprodukte der computerunterstützten Textedition

Hans Fix, University of Minnesota, Minneapolis

Als man im 19. Jahrhundert begann, die überlieferten aisl.Texte mehr oder weniger systematisch zu edieren, wurde in den Einleitungen dieser Texteditionen oft auch über die sprachliche Form des Textes gehandelt. Zu einer detaillierten Beschreibung kam es jedoch selten oder nie, denn im allgemeinen war das Ziel der Darstellung weniger, den vorgelegten handschriftlichen Text zu beschreiben, als seine Abweichungen von der imaginären Norm, seine Besonderheiten und Merkwürdigkeiten. Diese Praxis hat sich bis heute kaum geändert.

Wenn wir unsere Wörterbücher und Grammatiken betrachten, so gilt für die großen im Prinzip dasselbe: Das Interesse liegt kaum an der handschriftlichen Überlieferung, also an der sprachlichen Form, sondern vor allem an der Überlieferung, am Inhalt. Kurz, bei Noreen findet man alles, man findet nur nicht, wo man es findet; und das gilt auch für Fritzner und Cleasby, wo man (fast) alles findet, sich aber nicht darauf verlassen darf, wenn es um die sprachliche Form geht, um die Vollständigkeit der Paradigmen.

Wörterbücher und Grammatiken sind zu einzelnen Handschriften im 19. Jahrhundert schon selten und werden im 20. noch rarer, wenn man von denen absieht, die erst viele Jahrzehnte nach dem Start publiziert wurden. Aus diesen Gegebenheiten müssen wir - leider - ableiten, daß sprachwissenschaftliche Studien auf einer soliden und breiten Textbasis eigentlich kaum möglich sind, weil die notwendigen Vorarbeiten fehlen. Ich bin zwar nicht der Meinung, daß sich unser Bild vom Altisländischen radikal änderte, daß wir plötzlich eine völlig andere Normale brauchten, wenn wir mit Hilfe solcher Wörterbücher und Grammatiken zu einzelnen Handschriften die Überlieferung genauer studieren könnten. Ich bin jedoch überzeugt, daß es eine Menge kleinerer Korrekturen gäbe und daß wir über einiges sichere Aussagen machen könnten, was wir jetzt lediglich vermuten, erschließen

oder behaupten, daß wir womöglich auch von einigen lieb gewordenen Vorstellungen ein für alle Mal Abschied nehmen müßten.

Eine Reihe solcher Hilfsmittel in geeigneter lokaler(?) und zeitlicher Schichtung würde uns schließlich auch erlauben, begründete Aussagen über sprachliche oder auch nur über graphische Veränderungen im Laufe des Mittelalters zu machen.

Fast alle unsere Textausgaben erlauben das eigentlich nicht, weil sie ihre editorischen Prinzipien nicht im einzelnen klarlegen und genau beschreiben, was sie tun und was sie lassen; sie sind deshalb für sprachliche Studien ungeeignet. Es gibt allerdings ein paar bedeutende Ansätze der Textedition im 19. Jahrhundert, wie sie für sprachwissenschaftliche Zwecke notwendig sind, die man aber vor allem wohl aus Zeit- und Kostengründen leider nicht weiterverfolgt hat (K. Gíslason, Um frumpartar íslenzkrar túngu í fornöld, Kaupmannahöfn 1846; Ágrip af Noregs Konunga Sögum, udg. V. Dahlerup, Kjöbenhavn 1880 (STUAGNL)). Über mögliche andere Gründe soll hier nicht spekuliert werden.

An diese, wenn auch kurzlebige Tradition von Konráð Gíslason und Verner Dahlerup anzuknüpfen, erlaubt uns jetzt die computerunterstützte Edition und läßt uns Texte so edieren, wie sie es versucht hatten. Über den Sinn eines derartigen Publikationsunternehmens gibt es sicher unterschiedliche Meinungen, wenn man heute mit weniger Aufwand eine Handschrift fotografieren kann und quasi ein Original hat, als sich eine solche Edition erarbeiten läßt. Jedoch hat mich bislang noch niemand mehr überzeugen können als Konráð Gíslason, der 1846 bereits zwei Arten von Editionen forderte: eine handschriftengenaue für sprachliche Untersuchungen für den, der die Handschrift selber nicht hat, - davon ist oben die Rede - und eine textgemäß normalisierte für den "gemeinen Leser". Man kann sich sicher noch weitere Möglichkeiten vorstellen, sich auch fragen, ob diese textgemäß normalisierte Version den gemeinen Leser nicht überfordert, und denkt dann an normalisierte Texte für den Anfängerunterricht, für den isländischen Leser etc.etc.

Der Vorteil einer handschriftengenauen Edition gegenüber der Fotografie ist der des interpretierten Originals gegenüber dem Original. Vieles, was die Fotografie vielleicht überhaupt nicht, die Handschrift selbst womöglich erst nach langem Studium hergibt, wurde bereits systematisiert und kann gezeigt werden. Die Transliterationsprinzipien werden im einzelnen offengelegt, und darüber mag man sich dann auch streiten.

In welcher Form man eine solche Transliteration der Öffentlichkeit zugänglich macht, gedruckt, auf Mikrofiche oder maschinenlesbar, scheint mir vor allem eine Finanzfrage zu sein, denn daß solche Ausgaben wünschenswert sind, steht für mich außer Frage: Es ist eben einfacher, eine standardisierte Schrift zu lesen, selbst wenn sie viel diversifizierter ist als die konventionellen Buchschriften, als eine Handschrift. Daß die moderne Drucktechnik solche Ausgaben ermöglicht, und wie man es macht, wird die angekündigte Ausgabe der Möðruvallabók zeigen, daß es prinzipiell möglich ist, zeigen die Text in der Datenbank des Arnamagnäanischen Instituts. Das Wesentliche aus unserem Blickwinkel ist weniger das Buch im Faksimiledruck als der maschinenlesbare Text, bei dem man auf alle Einzelheiten zurückgreifen kann, soweit sie bei der Datenaufnahme berücksichtigt wurden. Eine solche Textversion wird der Ausgangspunkt für computerunterstützte sprachwissenschaftliche Studien sein. Die Datenbank des Arnamagnäanischen Instituts enthält mittlerweile zwei große und drei kleinere Transliterationen, nämlich

GkS 1157 fol. Grágás Konungsbók und AM 132 fol. Möðruvallabók

AM 334 fol. 92v–108 Staðarhólsbók (Járnsíða) sowie

AM 674a 4^{o} und AM 675 4^{o} (Elucidariusfragmente),

die Interessenten zur Verfügung gestellt werden können. Als Basis für diachrone Studien sind diese wenigen Handschriften vielleicht noch nicht hinreichend, es steht jedoch außer Zweifel, daß es sich bei ihnen um ganz bedeutende Denkmäler an Alter und Umfang handelt. Um dem Ziel sprachlicher Längsschnitte näher zu kommen, wäre es jetzt wichtig, weitere große und bedeutende Handschriften in

passendem zeitlichem Abstand zu dem bereits Vorhandenen aufzunehmen.
Die maschinenlesbaren Versionen dieser Handschriften folgen bislang unterschiedlichen Codiersystemen, die von den einzelnen Bearbeitern entwickelt wurden. Die Texte sind auch nicht mit gleicher Detailgenauigkeit transliteriert. Wir hoffen jedoch, daß wir vor allem durch diese Diskussion zu einem Standard kommen, dem alle Texte in der Datenbank folgen. Was der Einzelne damit dann auf seinem Computer macht, ob er den Standard beibehält oder nicht, bleibt natürlich seine Sache. Hier soll lediglich festgelegt werden, daß alle Texte in der Datenbank nach demselben System codiert sind, was die Annahme und Abgabe von Texten sehr vereinfachen würde.
Ich habe im Titel Wörterbücher und Grammatiken als Folgeprodukte der Textedition angesprochen, und ich glaube, der Wunsch nach beidem bedarf nach der obigen Lagebeschreibung keiner weiteren Rechtfertigung, wenn dies als Ziel von den Editoren bislang auch kaum verfolgt worden ist, vielleicht aus verständlichen Gründen: Nach jahrelanger Beschäftigung mit einem Text möchte man auch einmal etwas anderes machen, befürchtet, sich dem Spott der Zunft auszusetzen, man könne ja sonst nichts, und läßt so das Wissen um viele Einzelheiten verlorengehen. Selbst wenn das Produzieren einer Edition, einer Grammatik und eines Wörterbuches nach wie vor zu den philologischen Pflichtübungen gerechnet wird, dürfen sich - horribile dictu - in vieler Augen die Produkte nicht auf denselben Text beziehen.
In einem Wörterbuch nachzuschlagen, ob ein bestimmtes Lemma in einer Handschrift vorkommt, in einem bestimmten Text, in welcher Form, in welchem Kasus etc. geht eben viel einfacher und schneller als den Text zu lesen und hinterher nicht ganz sicher zu sein, ob man vielleicht doch etwas übersehen hat. Ein vollständiges Wörterbuch reduziert die Mühe der Belegsuche gewaltig, und das gilt natürlich auch für eine grammatische Darstellung, die das Material des Wörterbuchs, also der Handschrift, lediglich unter anderen Gesichtspunkten ordnet und interpretiert. Wenn wir Fragen nachgehen wie etwa, wie weit der Paradigmenwechsel bestimmter

Wörter zu einer bestimmten Zeit gediehen ist, ob sich die neuen Endungen der 1. Person Singular bereits durchgesetzt haben, wie die bestimmten suffigierten Artikel aussehen, mit welchen Zeichen bestimmte Phoneme dargestellt werden können, wie die Abfolge der Adverbiale im Satz ist, welche Valenz bestimmte Verben haben usw., so müßte uns die Textgrammatik eine Antwort in Bezug auf unseren Text liefern.

Wie lassen sich aus der Edition die Folgeprodukte gewinnen?

Es ist heute kein Problem mehr, aus einem maschinenlesbaren Text einen alphabetischen Index herzustellen; dafür gibt es konfektionierte Programme, die vielleicht mehr oder weniger geeignet sind, weil das Alphabet nicht vorgegeben werden kann oder anderes nicht genau so ist, wie man es gerne hätte, im Prinzip können aber alle einen Index produzieren. Dieser Index ist der Ausgangspunkt für alle Wörterbucharbeit, hat aber als KWIC-Index zudem einen Eigenwert, den man nicht unterschätzen sollte.

Als Ziel der Wörterbucharbeit zu Handschriften ist meines Erachtens ein vollständiger lemmatisierter Index das unabdingbare Minimum. Dazu können natürlich alle möglichen Ergänzungen gemacht werden, die die Sache verbessern und erweitern, Übersetzungen, Kollokationen, Frequenzen etc.etc. Vollständig muß der Index sein, damit man alles finden kann, was in der Handschrift steht, lemmatisiert, damit das auch ohne detaillierte Kenntnis der Schreibkonventionen der jeweiligen Handschrift möglich ist; denn wem fallen schon alle Möglichkeiten der Notation von /ǫ/ ein, wenn er Belege sucht.

Wegen dieser ungeregelten, manchmal vielleicht sogar wirren Graphie einer Handschrift, die sich durch eine überschaubare Anzahl von Regeln nicht normalisieren läßt, ist automatische Lemmatisierung, wie sie für die Gegenwartssprachen in diversen Projekten erarbeitet wurde, nur ansatzweise zu denken. Ich möchte jedoch nicht so weit gehen wie F. de Tollenaere vor 20 Jahren, der meinte, handschriftliche Texte seien für die automatische Lemmatisierung gänzlich ungeeignet,

es lohne die Mühe nicht. Ich glaube vielmehr, daß man durch Normalisierung des handschriftlichen Textes in textgerechter Weise eine Ebene herstellen kann, die als Basis für eine teilautomatische Lemmatisierung geeignet und hinreichend ist. Dabei ist von vornherein klar, daß es nicht um ein perfektes System gehen kann, sondern daß mit relativ wenigen und einfachen Umschreiberegeln der handschriftliche Text so verändert wird, daß er den Einträgen eines Arbeitswörterbuches (AWB) möglichst entspricht. Dieses AWB wird am Index dann vorbeigeführt, und die Einträge in beiden werden verglichen. Bei jeder Übereinstimmung erzeugt das Vergleichsprogramm einen oder mehrere Wörterbucheinträge, je nachdem wie viele Homographen das AWB enthält. Alle Wörterbucheinträge sind mit einer Mehrdeutigkeitsmarkierung versehen, die aus dem AWB stammt und die Korrektur erleichtern soll. Das AWB enthält ca. 11.500 Einträge, bestehend aus (1) Wortformen mit (2) Lemmanamen und (3) grammatischen Angaben, die bei den flektierenden Wortklassen den Maximalrahmen der paradigmatischen Mehrdeutigkeit beschreiben. Mit diesen Angaben kann ein Lemmatisierungsversuch gestartet werden. Das AWB ist anhand von Rechtstexten, nämlich Grágás und Jónsbók, erarbeitet worden und wurde vor kurzem auf die aisl. Elucidariusfragmente angewendet. Für mich war das Ergebnis frappant, denn wir konnten mit Hilfe des AWB auf Anhieb für 57% aller Wortformen des einen und für 67% des anderen automatisch Wörterbucheinträge erzeugen, obwohl sich aisl. Rechtstexte und sokratischer Dialog über Gott, Kirche und Welt inhaltlich doch recht fern stehen.

Die weitere Lemmatisierungsarbeit verläuft als Korrektur dieser automatisch lemmatisierten Indizes. Dabei müssen alle Einträge auf ihre Lemmazuweisung geprüft und, wenn paradigmatisch nicht eindeutig, vereindeutigt werden, soweit dies möglich ist. Wortformen, die beim AWB-Vergleich keinen Lemmatisierungsvorschlag erhielten, fehlen im AWB und müssen bestimmt werden. Sie kommen mit paradigmatischem Maximalrahmen außerdem ins AWB und stehen dann für weitere Lemmatisierungen mit zur Verfügung.

Nach Abschluß der Korrekturen wird das Material nach Lemmanamen alphabetisiert, innerhalb der einzelnen Lemmata nach grammatischen Kategorien. Die dann vorliegende Version sieht so aus, wie man es üblicherweise im Druck erwartet: nach dem Stichwort und der Wortklassenangabe folgen die Belege in Originalgraphie grammatisch sortiert, so daß der Nominativ vor dem Genitiv, der Singular vor dem Plural steht, daneben finden sich Häufigkeitsangaben. Teile des Materials werden in weiteren speziellen Verzeichnissen wie rückläufig alphabetisierte Wortformen, Wortformen nach fallender Frequenz etc. zusammengefaßt.

Solange das neue Textwörterbuch maschinenlesbar, d.h. dynamisch bleibt, ist es natürlich auch per Programm benutzbar und durchzusehen. Der Computer findet mit einem entsprechenden Suchprogramm beispielsweise alle Dative Sg.mask. auf -e der Substantive auf -r in unserem Text viel schneller, als man sie in der gedruckten statischen Version finden könnte, selbst wenn mit Hilfe des rückläufigen Wortformenverzeichnisses nur die Wörter betrachtet werden, die auf -e enden. Unter ihnen befinden sich zwar die gesuchten Wortformen, aber nicht alle auf -e sind auch die gesuchten Dative der Substantive auf -r. Das Suchprogramm wird das dynamische Wörterbuch jedoch genau nach diesen Kriterien durchkämmen. Der Nachteil eines maschinenlesbaren Wörterbuchs ist, daß es maschinengebunden ist, daß man ohne Programm nichts damit anfangen kann, während die gedruckte Version ihre Informationen jederzeit in der vorgegebenen Ordnung hergibt. Mühe hat man erst, wenn man Dinge wissen will, die nicht direkt ablesbar sind, denn das Buch läßt sich nicht verändern. Die herkömmlichen Wörterbücher ordnen alphabetisch vor- und allenfalls noch rückläufig, nach Häufigkeiten, nach Wortklassen, nach Sachgruppen und bieten ihr Material in dieser und nur in dieser Form, die sicher einen Teil der möglichen Fragestellungen beantwortet. Sie versagen jedoch, wenn man eine Kombination von Kriterien anlegen will, wie man es per Suchprogramm an ein entsprechend strukturiertes maschinenlesbares Wörterbuch kann. Das Suchprogramm zielt auf ein sehr spezielles Verzeichnis, kaum von allgemeinem Interesse und

daher zur Publikation nicht geeignet, jedoch von größter Bedeutung für den Benutzer mit einer sehr speziellen Fragestellung. Es wird wohl noch eine Weile dauern, bis maschinenlesbare Wörterbücher allgemein üblich werden; ob sie dann das gedruckte traditionelle Buch ersetzen können, möchte ich heute noch bezweifeln. Ob das wünschenswert wäre, ist eine ganz andere Frage. Die technischen Möglichkeiten dazu sind seit der Einführung der Kleincomputer sicher gegeben.

Als zweites Folgeprodukt der maschinenunterstützten Textedition ist die Grammatik angesprochen. Wir müssen hier zwischen dem graphemisch-phonemischen Teil, der Lautlehre, der vor allem in Zusammenhang mit der Transliteration erarbeitet werden wird, und den Teilen Morphologie und Syntax unterscheiden, die aus dem Wörterbuch erarbeitet werden. Der Computer leistet sehr wichtige Vorarbeiten bei der Erstellung des Graphem- und Allographeninventars, indem er die Umgebungen aller Zeichen dokumentiert. Die Umgebungsanalyse der Graphe ist für die Feststellung der Unabhängigkeit und damit für ihre Etablierung als Grapheme unerläßlich. Auf ein gesichertes Graphemsystem - und nur davon sollte zunächst geredet werden - kann mit den notwendigen Prämissen ein Phonemsystem gebaut werden. Minimalpaare auf graphischer Ebene, die Allographie demonstrieren, sind dem Wörterbuch leicht zu entnehmen; sie finden sich unter demselben Lemma mit derselben grammatischen Information.

Durch grammatische Sortierung kann man aus dem lemmatisierten Index schließlich die Grundlage für die Darstellung der Morphologie gewinnen; vgl. dazu methodisch W.Lenders/K.P.Wegera, Hrsg., Maschinelle Auswertung sprachhistorischer Quellen, Tübingen 1982 (Sprache und Information 3). Alle Substantive werden nach Genera zusammensortiert, dann die Endungen verglichen. Dabei müßten sich schnell die verschiedenen Klassen abzeichnen, die Belastung der einzelnen Klassen, der einzelnen Kasus, die Lücken in den Paradigmen ergeben. Es wäre interessant zu sehen, wie viele vollständige Paradigmen der Elucidarius etwa bietet, wie viele durch Berücksichtigung von Komposita oder anderer in dieselbe Klasse gehörende Wörter

aufgefüllt werden können und wie viele Lücken bei immerhin 8.000 und 12.000 Wörter langen Texten bleiben, welcher Art die Unterschiede zwischen älteren und jüngeren Fragmenten sind. Es scheint mir keine Frage zu sein, daß die Morphologie, wie Noreen sie etwa bietet, aus dem gesamten sprachlichen Material gewonnen werden kann und wenn nicht vollständig, dann doch mit großer Sicherheit rekonstruiert. Für die Sprachgeschichte wäre es aber nicht unwichtig, genau zu wissen, wie viel man bis zu welcher Zeit auch genau belegen kann, was einen größeren Zeitraum zur Belegbeschaffung braucht, was überhaupt nicht belegbar ist.
Durch syntaktische Uminterpretation der Wortarten und der grammatischen Informationen des lemmatisierten Index ist mit Hilfe der Stellenangaben ein syntaktisch analysierter Text erzeugbar. Auch das wird natürlich nicht fehlerfrei laufen, denn es ist klar, daß ein Nominativ zwar Subjekt oder Gleichsetzungsnominativ sein muß, ein Kasus obliquus kann jedoch Objekt, Attribut oder (Teil einer) Adverbiale sein. Eine syntaktische Konkordanz (vgl. K.Kossuth, Syntactic Concordances for Old Norse, in: H.Fix, Hrsg., Jenseits von Index und Konkordanz, Frankfurt usw. 1984 (Texte und Untersuchungen zur Germanistik und Skandinavistik Bd. 9), 155-186) wird uns Einsicht in Wortstellung, Valenzverhalten etc. gewähren können.

Eine neue Ausgabe aller aisl. Elucidariusfragmente mit einem vollständigen Wörterbuch und einer Grammatik ist sicher ein ambitioniertes Vorhaben, aber ein lohnendes und mit Computerunterstützung auch zeitlich machbares, wie mir scheint.