Martin-D. Gleßgen

Philologie und Sprachgeschichtsschreibung in der Romanistik: Die ‚informatische Wende'

1. Der philologische Zugang zur Sprachgeschichtsschreibung

1.1. Philologische Traditionen in der Romanistik

Der vorliegende Band macht die Verbindung zwischen Editionsphilologie und Sprachgeschichtsschreibung zum Thema. Die Philologie ist also nicht um ihrer selbst willen Gegenstand der Betrachtung, sondern in ihrem (potentiellen) Beitrag zur Sprachhistoriographie: Was bringt die Edition für die Sprachgeschichte und inwiefern formt sie sprachgeschichtliche Untersuchungen bereits im Vorfeld? Während diese Fragen im übrigen Band anhand spezifischer germanistischer Studien untersucht werden, sollte hier ein romanistisches Vergleichspaket geschnürt werden. Dies kann jedoch nur ganz impressionistisch versucht werden, da die Dichte und Vielfalt der philologischen Traditionen in der Romanistik sich dem synthetischen Zugriff entzieht.

Zunächst einmal verfügen in der Romania allein fünf große, sprachintern deutlich abgrenzbare Sprachräume über eine umfassende mittelalterliche Überlieferung: das Französische in Nordfrankreich, England (= Anglonormannisch) und – in deutlich geringerem Umfang – Norditalien (= ‚Franko-Italienisch'), das etwas später einsetzende, regional stark variante Italienische mit zahllosen Texten auch nichtliterarischer Ausrichtung (man denke an die 125.000 Briefe im Archiv von Francesco di Marco Datini), das relativ homogene Spanische mit seiner wichtigen wissenschaftlichen Literatur, der Flickenteppich des Okzitanischen, Gascognischen und Katalanischen (mit den Varietäten der Balearen und von Valencia), schließlich das insgesamt schlecht erschlossene Gallego-Portugiesische. Die übrigen Räume der Romania sind von viel geringerer Bedeutung (das Sardische kennt im Mittelalter nur Gebrauchsschriftlichkeit, das Rumänische scheint nur in der Onomastik auf, die Verschriftung des Frankoprovenzalischen, Bündnerromanischen oder Ladinischen ist kaum oder gar nicht entwickelt); doch es bleiben fünf große Konglomerate, die deutliche sprachinterne und textgeschichtliche Eigenheiten aufweisen.

Hinzu kommt eine große Vielfalt nationaler Erschließungstraditionen. Besonders in Italien, Frankreich und Spanien, aber auch in den mehrsprachigen Ländern Belgien und der Schweiz bestehen bedeutende philologische Traditionen in der Beschäftigung mit den jeweiligen Landessprachen. Dazu kommen alloglotte Editionsunternehmungen in Deutschland, den Niederlanden, Skandinavien, USA, Kanada oder Großbritannien. Wir müssen eine Größenordnung von etwa zehn nationalen Traditionssträngen annehmen, die sich mehr oder weniger intensiv den fünf großen Sprachgruppen wid-

men. Gewiss gibt es präferentielle Bindungen nach den Landessprachen, aber es kommt auch zu vielfältigen Querverbindungen; das Okzitanische etwa wird in den meisten Ländern mit romanistischer Tradition mehr oder weniger intensiv untersucht, was zur Überlagerung der jeweiligen Editionsmodelle führt.

Es ist ungemein schwierig, die zahlreichen nationalen und sprachgebundenen Traditionsstränge romanistischer Philologie herauszukristallisieren;[1] vielleicht ist es sogar unmöglich angesichts der zahlreichen Interferenzphänomene: Allein die Identifizierung möglicher Gliederungskriterien, die eine Strukturierung der Einzeleditionen erlauben könnten, wäre keineswegs trivial.[2] Eine ‚romanistische Eigenart' einer germanistischen Methodenvielfalt entgegenzustellen, ist gänzlich illusorisch, nicht nur in diesem Rahmen. Ich habe daher nicht versucht, den Entwicklungsgang ausgewählter Traditionen in der Romanistik nachzuzeichnen, sondern die wesentlichen aktuellen Grundgedanken über Fragestellungen und Perspektiven darzulegen, so wie sie sich mir als Romanist darstellen.[3] Vermutlich werden die meisten germanistischen Leser vieles finden, was ihnen keineswegs eigentümlich romanistisch erscheint; andererseits könnten verschiedene romanistische Leser sich möglicherweise nicht in der Kombination der vorgelegten Thesen wiederfinden. Die Fächergrenzen sind – zum Glück – in der Philologie recht offen.

1.2. Philologie vs. Sprachhistoriographie

Das Verhältnis zwischen Philologie und Sprachgeschichtsschreibung weist profunde Verwerfungen auf. Die Philologie thematisiert eine Reihe von Grundfragen, deren Behandlung beachtliche Energien verschlingt, die aber nur ganz mittelbar für die Sprachgeschichte relevant werden. Es sind dies die unvermeidlichen Überlegungen über die zentralen Editionsentscheidungen (diplomatische Edition von Einzelmanuskripten, synoptische Edition, kritische Edition), die je nach Textsorte und Überlieferungslage in die unterschiedlichsten Kompromissformen münden; es sind weiterhin Entscheidungen über die präzise Form des Variantenapparats und der Kommentarformen: Wie weit geht die sprachwissenschaftliche ‚Primärkommentierung'? Das heißt konkret: Welchen Umfang haben Glossare, Indizes, Untersuchungen zum Sprachstand? Es sind schließlich die oft intensiven Studien zur Datierung und Lokalisierung der jeweiligen Texte und Manuskripte, die erst eine Verankerung der jeweiligen sprachlichen Zeugnisse im historischen Diasystem ermöglichen.

---

[1] Vgl. den Versuch von Frédéric Duval (Hrsg.): Pratiques philologiques en Europe. Actes de la journée d'études à l'Ecole Nationale des Chartes (23 sept. 2005). Paris [im Druck] (Etudes et rencontres); vgl. auch meinen mit Franz Lebsanft herausgegebenen Band: Alte und Neue Philologie. Tübingen 1997 (Beihefte zu editio. 8).

[2] Als ein hilfreicher Gliederungsparameter erwies sich – um nur ein Beispiel zu nennen – der Ansatz, die Methodik der editionsbegleitenden Glossare einer strukturierenden Analyse zu unterwerfen; vgl. die Studie von Sylvie Korfanty anhand der Editionen französischer Texte des 16. Jahrhunderts: Lexicographie et glossographie du français du XVIème siècle. Prolégomènes à un dictionnaire du français préclassique. Lille, Atelier National de Reproduction des Thèses, 1999 [3 Microfiches].

[3] Dieser Entscheidung, eine persönliche Synthese zu versuchen, ist der kontinuierliche Verweis auf meine entsprechenden thematischen Untersuchungen geschuldet, was die Argumentation an dieser Stelle straffen kann.

Erst danach stellt sich die Frage möglicher weitergehender sprachwissenschaftlicher Analysen auf der Grundlage der edierten Texte und es eröffnet sich damit der Blick auf die Sprachgeschichte. Diese wiederum legt ganz andere Raster an: Sie untersucht die verschiedenen sprachlichen Segmente – Graphematik, Phonologie, Morphologie, Syntax, Lexik und Onomastik – in den überlieferten Textsorten über die Zeiten hinweg. Der Text ist nicht mehr Selbstzweck, sondern Steinbruch für ganz spezifische Einzelfragen.

Der Grundkontrast zwischen Text und Sprache beherrscht seit dem 19. Jahrhundert das Wechselspiel zwischen Editionsphilologie und Sprachgeschichtsschreibung. Der Philologe kann sinnvollerweise immer nur die Bedeutung des ihm vorliegenden Textes anhand der in ihm verwendeten Sprachzeichen untersuchen; die im eigentlichen Sinne diachrone Arbeit ist daher für ihn unorganisch.[4] Andererseits ergeben sich Innovationen für die Sprachgeschichtsschreibung nur an der Zündstelle zwischen Einzeltext und anderweitig bekannten sprachlichen Daten, was zu einer punktuellen Synthese der zwei unterschiedlichen Sichten – der auf den Text und der auf die Sprache – zwingt.

Strukturell gesehen bleibt das Wechselspiel unausgeschöpft, wie es die historische Lexikographie vielleicht am einfachsten verdeutlicht: Das Editionsglossar zu einem galloromanischen Text sollte vernünftigerweise alle Scharnierbelege gegenüber der Lexikographie (bes. *Französisches Etymologisches Wörterbuch* [FEW] und *Trésor de la langue française du XIXe et XXe siècle* [TLF]) vermerken; was aber nützt es, wenn eine bestimmte Form und Bedeutung in fünf verschiedenen Editionen als „neuer Erstbeleg gegenüber dem FEW" vermerkt wird? Wo wird diese Information in zugänglicher Weise vermerkt? Das Studium des Einzeltextes erbringt also neue Resultate in der Kenntnis der Sprachgeschichte, was aber nur anhand der bestehenden Überblicksrepertorien für den Einzelnen erkennbar sein kann; die Repertorien jedoch werden im Augenblick der Korrektur überholt.

Die Problematik zeigt sich in noch größerer Schärfe in allen übrigen Bereichen der Sprache, für die wir in der Romanistik traditionell nicht über ähnlich gute Repertorien verfügen wie in der Lexik; keine historische Grammatik, aber auch keine historische Phonologie oder Onomastik kann für die Galloromania dem FEW das Wasser reichen. Das erschwert dann zusätzlich das Hin und Her zwischen dem Text und der Sprache.

2. Die ‚informatische Wende' in der Editionsphilologie

2.1. Neue Perspektiven

Die jüngeren, rasanten Entwicklungen in der Informatik eröffnen nun radikal neue Perspektiven im Spannungsfeld zwischen Editionsphilologie und Sprachgeschichte.

[4] Diesen Gedankengang verfolgt stringent Gerhard Ernst: Lexikalische Analyse historischer Texte und semantische Theorie am Beispiel nonstandardsprachlicher französischer Texte des 17. und 18. Jahrhunderts. In: Historische Semantik in den romanischen Sprachen. Hrsg. von Franz Lebsanft und Martin-D. Gleßgen. Tübingen 2004 (Linguistische Arbeiten. 483), S. 153–161.

In ähnlicher Weise wie die ‚kognitive Wende' der siebziger und achtziger Jahre ein neues Fundament für die sprachwissenschaftliche Deutung gelegt hat, das zugleich den Strukturalismus sinnvoll integrieren konnte, schafft die ‚informatische Wende' eine gänzlich neue Situation in der Sprachgeschichtsschreibung. Bei Berücksichtigung einiger Grundanforderungen kann jetzt jeder Einzeltext von der Edition ab dauerhaft Teil eines weltweit abfragbaren Textdatenbestandes werden, wodurch der grundsätzliche Kontrast zwischen individuellem, konkretem Diskurs und abstrakter, nach funktionalen Segmenten organisierter Sprache entschärft wird.

Der einzelne Text wird in den allmählich entstehenden großen Korpora zum Vertreter einer bestimmten Textsorte, gebunden an die übrigen diasystematischen Parameter von Zeit, Raum und sozialem Prestige. Er steht neben anderen, ähnlichen Texten und kann mit diesen quantifiziert werden. Es ist damit auch möglich, unterschiedliche Manuskripte einer Texttradition nebeneinander zu stellen und sie zu befragen, was bisher selbst im Rahmen einer synoptischen Edition sehr mühsam war. Weiterhin können sich nun die Untersuchungen zu Sprachwandel und Sprachgeschichte unmittelbar auf die überlieferten Textzeugen stützen, ohne den zwingenden Umweg über Einzelglossare oder graphematische und grammatische Einzelanalysen.

Die Rolle der Textsorten erfährt in einem korpuslinguistischen Ansatz eine besondere Aufmerksamkeit, da jede lexikalische oder syntaktische Sprachform zunächst einmal in ihrer Bindung an die entsprechende Gattung zu untersuchen ist. Aus sprachtheoretischen Gründen ist auch dies ein perspektivenreicher Aspekt der Computerphilologie, da der Weg sprachlicher Veränderungen im Diasystem und in der sprachlichen Konfiguration (= im ‚Sprachsystem') von der einzelnen Sprachäußerung über die aus dieser hervorgehende bzw. diese prägende Textsorte führt.[5]

2.2. Anforderungen an die historische Korpuslinguistik

Die Perspektiven einer historischen Computerlinguistik haben den Nachteil, dass sie eine ganze Reihe von neuen Zwängen schaffen, die nicht ohne den entsprechenden Aufwand zu umschiffen sind. Eine philologisch tragfähige historische Korpuslinguistik muss bestimmte Anforderungen erfüllen:

1. Die Textdaten müssen für eine Langzeitsicherung angelegt sein; das bedeutet eine Kodierung in XML und damit sinnvollerweise auch die Verwendung von anwendungsneutralen Texteditoren oder Redaktionssystemen. Unicode wird inzwischen von nahezu allen Systemen unterstützt, und auch die Vorschläge der TEI sind als Referenz bekannt. Für die – unumgängliche – Beschreibung der Tagsets bietet das – erst 2001 entwickelte – (XML-basierte) Schema mehr Möglichkeiten als die DTD. Trotz einer gewissen Schwerfälligkeit ist dies ein Bereich, wo heute Standards herrschen, die alle nötigen Anforderungen für Langzeitsicherung und Offenheit der Auswertung und des Exports garantieren. Der Export in Druckform oder in eine Browserdarstellung ist augenblicklich jedoch noch etwas mühsam, und die eigentliche sprachwissenschaftli-

[5] Vgl. Martin-D. Gleßgen: Diskurstraditionen zwischen pragmatischen Regeln und sprachlichen Varietäten. In: Historische Pragmatik und historische Varietätenlinguistik in den romanischen Sprachen. Hrsg. von Angela Schrott und Harald Völker. Göttingen 2005, S. 207–228.

che Auswertung wirft gewichtige Probleme auf, da sie an bestimmte Programme gebunden ist.

2. Die verwendeten Programme sollten so transparent, zugänglich und dauerhaft wie möglich angelegt sein. In diesem Bereich sind wir leider noch sehr weit von möglichen Referenzen entfernt. Es existiert eine große Vielzahl korpuslinguistischer Instrumente für unterschiedliche Zwecke und mit unterschiedlichen Qualitäten. Für die Wissenschaft sind langfristig allein Open-Source-Produkte tragfähig (= Transparenz), die frei und kostenlos zugänglich sind (= Verfügbarkeit). Die Nachteile kommerzieller Produkte liegen meist nicht nur im Preis, sondern auch in der Schwierigkeit des Datenexports. Noch unzugänglicher und daher problematisch sind die Programmpakete der großen Sprachinstitutionen, etwa COSMAS (DUDEN) – technisch dabei sehr hilfreich für Kollokationen – oder das Programm STELLA des französischen Pendants (TLF, vgl. supra 1.2.), das vielfältige Suchgruppen ermöglicht.

In der eigenen Praxis verwende ich als Texteditor – und zum Teil als Skriptsprache – Tustep (frei zugänglich, Open-Source angekündigt), als XML-Editor die (kostenlose) Home-Version des XML-Spy und arbeite mich zurzeit in die XML-Produkte zur sprachwissenschaftlichen Abfrage ein (X-Path, XSLT und bes. X-Query). Der allseits gelobte Editor E-Macs ist wie die Skriptsprachen Perl oder Python relativ zeitaufwendig im Erlernen. Für die Lemmatisierung und Datenbankverwaltung habe ich kein Instrument gefunden, das allen Anforderungen einer komplexen mittelalterlichen Textedition gerecht wird, was mich in den mühsamen Weg einer Eigenprogrammierung führte (s.u. 3.). Für das morphologische Tagging (alt-)französischer Texte verwende ich den Tree-Tagger (kostenlos unter LINUX).

Schon dieser kurze, ganz persönliche Überblick lässt erkennen, wie schwierig es ist, sich zu orientieren, zudem für nicht mathematisch Geschulte und oft – wie in meinem Fall – auch nicht besonders mathematisch Begabte. Die Probleme sind jedoch nicht individueller, sondern struktureller Natur, und ihre Klärung wird uns noch einige Zeit beschäftigen.

3. Die informatische Aufbereitung der Textdaten muss mit einer philologischen Aufbereitung und sprachwissenschaftlichen Primärerschließung klassischer Ausrichtung einhergehen. Die Textgrundlage ist die Referenz für alle sprachwissenschaftlichen Abfragen und bedingt auch die möglichen Funktionalitäten der jeweiligen Analysetools. Die Texterstellung ist nun bei älteren Texten – im Gegensatz zu aktuellen Korpora – nicht trivial.

Ein erster zentraler Punkt ist die Wortsegmentierung: Eine altfranzösische Form *alabe* im Manuskript sollte in der Form *à-l'abé* transkribiert werden, eine Form *ala be* in der Form *à-l'a_bé*; das erlaubt es, jeweils die mittelalterliche Trennung zu rekonstruieren, aber auch zugleich eine automatische Wortsegmentierung vorzunehmen. Das hier angewandte Prinzip der doppelten Kodierung, das eine mittelalterliche und moderne Lesung ermöglicht, ist auch für die Groß- und Kleinschreibung, für Satzzeichen und Textformat eine sinnvolle Lösung; hinzu kommt die Korrektur von *Lapsus calami* (*conite* ist zu korrigieren in *comte* [mit der entsprechenden Anmerkung]), da das Wort oder Morphem als geteiltes, intersubjektives Wissen Gegenstand der Sprachwissenschaft ist; auch deutende Kommentare sind neben den Emendationen

punktuell erforderlich. Allzu oft vernachlässigt wird schließlich die Textsegmentierung, die eine anderweitig verfügbare oder neu als gültig eingeführte Referenznummerierung verwendet.

Die informatische Aufbereitung vermindert in keinem Fall den Aufwand der Sichtung einer Manuskripttradition. Es ist durchaus möglich – wenngleich sehr arbeitsintensiv –, im Sinne einer synoptischen Edition acht Handschriften eines Textes nebeneinanderzustellen wie im Fall des *Chevalier de la charrette*;[6] die 600 Mss. der *Divina Commedia* dagegen übersteigen alle Möglichkeiten der Mittelalterphilologie.

Die klassischen Anforderungen der Verortung im Diasystem stellen sich, wie angedeutet, eher noch in verschärftem Ausmaß als in der traditionellen Philologie, da es darum geht, größere Korpora auf einmal zu befragen, wobei die einzelnen Textzeugen gegeneinander abgegrenzt werden müssen. Erforderlich sind die Datierung und Lokalisierung von Text und Handschrift, die Identifizierung der Textsorte – soweit dies beim aktuellen Forschungsstand möglich ist – und etwaiger (fremdsprachlicher, in der Romania primär lateinischer) Vorlagen sowie schließlich des Entstehungs-‚Ortes', also der Kanzlei oder des Skriptoriums, in dem die Handschrift vermutlich entstand. Letzteres erlaubt zugleich eine diastratische Zuordnung, während die diaphasische Bindung an der Textsorte ansetzt.[7]

4. Ein informatisch erfasster Text, der die drei genannten Kriterien erfüllt, kann gefahrlos in eine größere Textdatenbank eingefügt werden. Sollte er zugleich individuell ediert werden (in Druckform oder in Browserdarstellung), sind weitere Schritte der klassischen philologischen Arbeit erforderlich: Die Analyse des graphematischen und des morphologischen Systems, ein Glossar sowie ein Register der Orts- und Personennamen, das idealerweise in einer lexikologischen (und damit etymologischen) Struktur angelegt ist. Diese Primärerschließung beruht letztlich auf einer Gruppenbildung sprachlicher Eigenschaften in den verschiedenen Bereichen (Graphematik, Morphologie, Lexik, Onomastik).

Eine solche Quellenerschließung eröffnet ein umfassendes interpretatives Potential: Die Quantifizierung von Sprachdaten wird im Diasystem der Schrift möglich, Sprachwandel und Sprachausbau sind nach den Parametern des Diasystems (Zeit, Raum, Gesellschaftsgruppe [= Entstehungs-‚Ort'], Kontextbindung [vgl. Textsorte]) nachweisbar. Der entsprechende Aufwand ist allerdings nicht unerheblich.

2.3. Philologie vs. Korpuslinguistik in der Romanistik

Die beschriebene Entwicklung hat zunächst einmal keine romanistischen, nicht einmal genuin sprachwissenschaftliche Wurzeln; sie stammt aus der Welt der angewandten Informatik. Es ist jedoch nicht nur wichtig, welche informatischen Möglichkeiten

[6] Vgl. Cinzia Pignatelli: L'archive du projet «Charrette». Huit manuscrits prêts à se livrer. In: Ancien et moyen français sur le web. Hrsg. von Pierre Kunstmann, France Martineau und Danielle Forget. Ottawa 2003, S. 203–220.

[7] Vgl. Martin-D. Gleßgen: Vergleichende oder einzelsprachliche historische Textwissenschaft. In: Was kann eine vergleichende romanische Sprachwissenschaft heute (noch) leisten? Hrsg. von Wolfgang Dahmen et al. Tübingen 2006 (Romanistisches Kolloquium. 20. Tübinger Beiträge zur Linguistik. 491), S. 319–340.

bestehen, sondern auch, wie diese genutzt werden. Hierin nun haben die einzelnen Philologien neuerlich ihre Traditionen, die auf den älteren methodischen Errungenschaften beruhen. Es ist zum Beispiel kein Zufall, sondern Traditionswerk, dass die große altitalienische Datenbank des ‚Tesoro della lingua italiana delle Origini' (TLIO) auf einer exemplarischen philologischen Grundlage beruht.

In den Einzelphilologien erweisen sich auch die wirklich realisierbaren Perspektiven und die methodischen Probleme, die sich im Spannungsfeld von Philologie, Sprachgeschichte und Korpuslinguistik ergeben. Nehmen wir den Fall des Französischen: Eine inhaltlich gute Datenbank zum älteren Französisch ist die ‚Base du Dictionnaire du Moyen Français' (DMF), die in Nancy auf der informatischen Grundlage von ‚Frantext' (der Datenbank des TLF) erstellt wird. Da diese Datenbank schon relativ früh – in den achtziger Jahren – begonnen wurde, beruht das eigentliche Wörterbuch nicht in allen Einträgen direkt auf der Textdatenbank, was der verantwortliche Programmierer (Gilles Souvay) durchaus geschickt durch semiautomatische Routinen der Lemmatisierung auszugleichen sucht. Aber das Grundproblem des Verhältnisses zwischen Textdatenbank und sprachhistorischer Auswertung (hier im Wörterbuch) wird in exemplarischer Weise deutlich.

Ein anderes Problem zeigt der bereits zitierte Fall des *Chevalier de la charrette*. Wir verfügen über eine Computeredition der acht Manuskripte, was als sprachhistorischer Glücksfall gewertet werden könnte; eine philologische Prüfung ergab jedoch sehr schnell, dass die Transkriptionen sehr fehlerbehaftet waren und vor jeder wirklich skriptologischen Auswertung korrigiert werden mussten.[8] Da das Rezensionswesen informatische Editionen (bisher) nicht erfasst, sind solche Erscheinungen leider der Normalfall und nicht die Ausnahme in der Computerphilologie.

Die aktuellen Bestrebungen in der Alt-Französistik gehen hin zu einer wechselseitigen Annäherung der unterschiedlichen, an verschiedenen Stellen, zu verschiedenen Zeiten, mit unterschiedlichen Absichten erstellten Textkorpora.[9] Im – sprachhistorischen – Idealfall sollten die Korpora auf der (diplomatischen) Edition von Einzelmanuskripten beruhen, und die jeweiligen sprachwissenschaftlichen Auswertungen sollten in der Textdatenbank selbst ihren Niederschlag finden (also Eintrag der Lemmatisierung, graphematischer und grammematischer Identifizierungen).

2.4. Bilanz

Der aktuelle Stand der Dinge erlaubt damit folgende Thesen, die in ihrer Abstraktheit nichts Romanistisches haben:

1. Die Wissenschaftszweige der Philologie und der Sprachgeschichtsschreibung sind aufeinander angewiesen, repräsentieren jedoch zwei radikal unterschiedliche Sichtweisen auf die historische Realität.

---

[8] Vgl. nochmals Pignatelli 2003 (Anm. 6), S. 204.

[9] Vgl. den Band Le Nouveau Corpus d'Amsterdam. Hrsg. von Pierre Kunstmann und Achim Stein. Stuttgart [im Druck] (Beihefte der Zeitschrift für französische Sprache und Literatur).

2. Die Möglichkeiten der Computerlinguistik erlauben eine strukturelle Annäherung der beiden Gebiete, die zu einer wirklichen Synthese führen kann.
3. Der Aufwand, um die Computerlinguistik in Philologie und Sprachhistoriographie einzuführen, ist erheblich, sowohl in der zu investierenden Zeit wie in der Notwendigkeit, neue Konzepte zu entwickeln, die auch in die universitäre Lehre Eingang finden können.

Vor diesem Hintergrund versuche ich seit 1998 eine konkrete Arbeitsroutine zu entwickeln, die von der Textedition zur sprachhistorischen Analyse führt.[10] Das Projekt hat exemplarischen Charakter und soll nicht zuletzt zur Klärung der methodischen Möglichkeiten und Kosten führen; es hat allerdings auch einen definierten empirischen Kern mit konkreten sprachhistorischen Fragestellungen. In der Folge soll der aktuelle Stand der Überlegungen und Programmierung in aller Kürze vorgestellt werden, um die Betrachtungen der ersten beiden Kapitel an einem romanistischen Beispiel zu konkretisieren.

Die Ziele des Programmpakets Phoenix sind es, ältere romanische Texte nach den oben dargelegten Kriterien zu edieren, sie philologisch zu erschließen und für eine weitergehende sprachhistorische Analyse aufzubereiten.[11]

3. Exemplum: Phoenix

3.1. Edition und Datenkodierung

Das konkrete philologische Projekt, im Rahmen dessen Phoenix entwickelt wurde, ist die Edition und sprachwissenschaftliche Aufbereitung der *Plus anciens documents linguistiques de la France*,[12] eine umfangreiche Sammlung altfranzösischer Originalurkunden aus dem 13. Jahrhundert. Die Edition beruht auf einer anwendungsneutral kodierten Transkription der Manuskripte in XML; der jeweilige Urkundentext wird von einem ‚Tableau analytique' begleitet, das die diasystematische Verortung und die traditionelle philologische Beschreibung enthält.

---

[10] Vgl. Martin-D. Gleßgen: Das altfranzösische Geschäftsschrifttum in Oberlothringen: Quellenlage und Deutungsansätze. In: Skripta, Schreiblandschaften und Standardisierungstendenzen. Beiträge zum Kolloquium vom 16. bis 18. September 1998 in Trier. Hrsg. von Kurt Gärtner, Günter Holtus, Andrea Rapp und Harald Völker, Trier 2001, S. 257–294.

[11] Die Entwicklung von Phoenix erfolgte in Zusammenarbeit zunächst mit Matthias Kopp (Tübingen), dann auch mit Matthias Osthof (Tübingen und Zürich). Alle Entscheidungen über die Struktur der Programme entstanden in gemeinsamen Diskussionen. Das Lemmatisierungstool wurde von Matthias Kopp, das Datenbankmodell und das Schema mit den entsprechenden Prüfungsinstanzen von Matthias Osthof programmiert.

[12] Das Projekt wird in Zusammenarbeit mit der Ecole Nationale des Chartes (Françoise Vielliard, Olivier Guyotjeannin) verfolgt und steht in Zürich im Rahmen des Nationalen Forschungsschwerpunkts „Medialität (Medienwandel, Medienwechsel, Medienwissen)"; vgl. die Homepage NCCR Mediality, <http://www.mediality.ch/>.

Hier das Beispiel einer (verkürzten) Urkunde des Korpus[13]:

002
1234 (25 mars-31 décembre) ou **1235** (1er janvier-24 mars)

Type de document : **charte : acensement de terres**
Objet : *L'abbé et le chapitre de Salival acensent à Wirrion et Houillon treize journaux de terre au finage de Juvelize contre un cens de treize deniers et deux hémines de grain ; les conditions de l'acencement sont très contraignantes pour les paysans.*

Auteur : non annoncé
Disposant : abbaye de Salival
Sceau : disposant
Bénéficiaire : disposant [la rédaction de la charte avantage surtout le chapitre]
Autres acteurs : Wirrion et Houillon, paysans de Juvelize
Rédacteur : **scriptorium de l'abbaye de Salival** [les paysans ne pouvaient pas disposer d'un scribe]

Parchemin jadis scellé sur simple queue ; 58 x 141
AD MM H 1244, fonds de l'abbaye de Salival

1 Conue chose soit a-toz **2** q*ue* li abes *et* li chapitles de Salinvas · at laissé a Wirion / *et* Huillon, les dous freres de Gev*er*lise, les anfanz Bertran Bacheler, **3** ·XIII· jor/nas de t*er*re treisse · en la fin de Gev*er*lise · *et* a lor oirs · **4** p*ar*mi ·XIII· d*eniers* de cens · *et* / ·II· himas de blef · l'un d'avoine · l'autre de froment · **5** *et* s'il ne paievent a jor // nomei a la feste sent Remi · a Giv*er*lise, en la maison de Salinvas · q*ue* l'on se tan/roit a la terre · *et* ce q*ue* sus averoit ·
6 (…)
10 Ci at mis li abes et li convenz de Salinvas son sael · en tesmoig/nage de verité · **11** l'an q*ue* li miliaires corroit p*ar* ·M· *et* CC· *et* XXXIIII· anz ·

In der XML-Kodierung stellt sich dieser Ausschnitt – leicht vereinfacht – folgendermaßen dar:

```
<id>555550002</id>
<zitf>002</zitf>

<an>
<nom>002</nom>
<d>1234 (25 mars-31 décembre) ou 1235 (1er janvier- 24 mars)</d>
<d0>1234/09/25</d0>
<type>charte: acensement de terres</type>
<loc>Lorraine ducale</loc>
```

[13] Vgl. ausführlicher Martin-D. Gleßgen: Editorische, lexikologische und graphematische Erschließung altfranzösischer Urkundentexte mit Hilfe von TUSTEP. Stand der Arbeiten. In: Drittes Trierer Urkundensprachekolloquium (20.–22. Juni 2001). Hrsg. von Kurt Gärtner und Günter Holtus. Trier 2005, S. 91–107, hier S. 92–100.

```
<loc0>-</loc0>
<soc>Clergé régulier</soc>
<soc0>-</soc0>
<r>L'abbé et le chapitre de Salival acensent à Wirrion et Houillon treize journaux de terre au finage de Juvelize contre un cens de treize deniers et deux hémines de grain.</r>
<aut>non annoncé</aut>
<disp>abbaye de Salival</disp>
<s>disposant</s>
<b>disposant</b>
<act>Wirrion et Houillon, paysans de Juvelize</act>
<rd>scriptorium de l'abbaye de Salival [rdp: néant; com: soustraitance?]</rd>
<rd0>AbbSalival</rd0>
<f>Parchemin jadis scellé sur simple queue; 58x141</f>
<l>AD MM H 1244, fonds de l'abbaye de Salival</l>
</an>

<txt>
<div n="1"> <maj>C</maj>onue chose soit a-toz</div>
<div n="2"> q<abr>ue</abr> li abes <abr>et</abr> li chapitles de Salinvas /. at laissié a Wirion
<zw/> <abr>et</abr> Huillon, les dous freres de Gev<abr>er</abr>lise, les anfanz Bertran Bacheler,</div>
<div n="3"> /.XIII/. jor<zwt/>nas de t<abr>er</abr>re treisse,//. en la fin de Gev<abr>er</abr>lise /. <abr>et</abr> a lor oirs,//.</div>
<div n="4"> p<abr>ar</abr>mi /.XIII/. d<abr>eniers</abr> de cens /. <abr>et</abr>
<zw/> /.II/. himas de blef,//. l'un d'avoine,//. l'autre de froment;//.</div> (...)
<par/>
<div n="6"> (...)
<par/>
<div n="10"> <maj>C</maj>i at mis li abes <abr>et</abr> li covenz de Salinvas son sael,//. en tesmoig<zwt/>nage de verité,//. </div>
<div n="11"> l'an q<abr>ue</abr> li miliaires corroit p<abr>ar</abr> /.M/. <abr>et</abr> CC/. <abr>et</abr> XXXIIII/. anz.//.</div>
</txt>

</gl>
```

Die Vorteile dieser – an den Prinzipien der TEI (*Text Encoding Initiative*) orientierten – Kodierung liegen nicht zuletzt in einer relativ offenen Darstellbarkeit: Der Text kann automatisch in Druckform oder in eine Browserdarstellung überführt werden, wobei zudem Wahlmöglichkeiten für den Nutzer bestehen (diplomatische oder interpretative Edition). Die Konsistenz und Korrektheit der Kodierung wird durch ein XML-Schema geprüft und garantiert.

Für eine sprachwissenschaftliche Bearbeitung bietet diese Kodierung eine ideale Grundlage, da der Text sorgfältig segmentiert ist und jede einzelne Sprachform im Diasystem verortet wurde (die diasystematischen Parameter des Raumes [<loc> im Tagset] und des Sozialprestiges [<soc>] erscheinen in der obigen Druckversion nicht,

sind aber in der Datenbank gegenwärtig). Die philologische Sicherheit und Interpretierbarkeit ist durch eine solche Edition in idealer Weise gewährleistet, was den zusätzlichen Aufwand der nötigen informatischen Instrumente rechtfertigen kann.

3.2. Sprachwissenschaftliche Auswertung

Phoenix besteht aktuell aus zwei Programmen zur sprachwissenschaftlichen Anreicherung der Daten. Zunächst erlaubt ein Lemmatisierungstool eine Bildung lexikalischer, aber auch morphologischer oder graphematischer Gruppen: Einzelwörter können unter ein Lemma zusammengefasst werden oder Wörter, die eine bestimmte graphematische Eigenschaft aufweisen, mit einem entsprechenden Eintrag versehen werden (z.B. *Conue chose soit a-toz que li <wn n="328" lex=abbé> abes </wn> et li chapitles de Salinvas*).[14] Diese Primäranreicherung erfolgt zunächst ganz analog zur traditionellen – ‚händischen' – Textanalyse; sie kann allerdings auch ggf. auf zusätzliche Hilfsmittel wie Formenlexika und trainierte Tagger zurückgreifen.[15] Entscheidend ist, dass der Grundtext auch nach einer einmal erfolgten Anreicherung noch korrigiert werden kann, was die nötige Flexibilität garantiert.

Das zweite Programm erlaubt die Erstellung einer lexikologischen, onomastischen, graphematischen sowie morphologischen Datenbank. Die bereits attributierten Textdaten können aufgerufen und nach einem stringenten Modell detailliert beschrieben werden. Diese Datenbank wird zunächst die Erstellung der Glossare und Namensglossare für die einzelnen Urkundenbände ermöglichen, wobei einmal eingeführte Definitionen jeweils in neue Bände übernommen werden können. Parallel dazu führt sie allmählich zu einem Supplementwörterbuch der altfranzösischen Urkundensprache und zu einem ersten onomastischen Inventar des älteren Französisch.[16]

3.3. Ausblick

Die sprachwissenschaftlich angereicherten Textdaten können in einem weiteren Schritt für unterschiedliche Analysen verwendet werden, für die graphematische Normbildung ebenso wie für die Analyse von Textmodellen.[17] Dazu sind allerdings zusätzliche Abfrageinstrumente nötig, für die ich auf die Serie der genannten XML-Tools zurückgreifen möchte. Während die beiden Tools von Phoenix sicherlich auch für andere philologische Projekte interessant sein können, plane ich in diesen weiter-

---

[14] Vgl. Martin-D. Gleßgen, Matthias Kopp: Linguistic annotation of texts in non-standardized languages. The program procedures of the tool Phoenix. In: Romanistische Korpuslinguistik II: Korpora und diachrone Sprachwissenschaft / Romance Corpus Linguistics II: Corpora and Diachronic Linguistics. Hrsg. von Claus D. Pusch, Johannes Kabatek und Wolfgang Raible. Tübingen 2005 (ScriptOralia. 130), S. 147–154.

[15] Vgl. Achim Stein, Martin-D. Gleßgen: Resources and Tools for Analyzing Old French Texts. In: Romanistische Korpuslinguistik II (Anm. 14), S. 135–145.

[16] Vgl. Martin-D. Gleßgen: L'élaboration philologique et l'étude lexicologique des Plus anciens documents linguistiques de la France à l'aide de l'informatique. In: Frédéric Godefroy. Actes du Xe colloque international sur le moyen français. Hrsg. von Frédéric Duval. Paris 2003, S. 371–386.

[17] Vgl. Martin-D. Gleßgen: Les 'lieux d'écriture' et leur identification dans les documents lorrains du XIIIe siècle. In: Revue de Linguistique Romane [im Druck].

führenden Bereichen keine Standard-Programmierungen. Hier bietet die Computerlinguistik Anwendungsmodelle, die zwar für Philologen zum augenblicklichen Zeitpunkt noch schwer zugänglich sind, dies aber zweifellos nicht bleiben werden. Die Entwicklung neuer Standards erfolgt zurzeit kaum irgendwo mit solch rasanter Geschwindigkeit wie im Bereich von XML.[18]

Für die Sprachgeschichtsschreibung entstehen über den unbestreitbaren Aufwand bemerkenswerte Möglichkeiten: Etwa die Perspektive, ein Skriptoriennetz für das Altfranzösische zu identifizieren und an Einzeltexten festzumachen, die Quantifizierung von Erscheinungen in den verschiedenen Bereichen der Sprache oder der Nachweis des Sprachwandels nach Maßgabe der Textsorten und der Parameter des Diasystems. Der Einzeltext wird Teil eines größeren Korpus, das einen Bogen zur historischen Sprache spannt.

Diese Entwicklungen und Perspektiven haben, wie gesagt, keine unmittelbar offensichtliche romanistische Bindung. Sie werden gleichwohl durch den romanistischen Kontext und durch die hier angesiedelten dichten philologischen Traditionen geprägt und erhalten dadurch ihre inhaltliche Verankerung. Für das Spannungsfeld von Philologie und Sprachgeschichtsschreibung bietet die ‚informatische Wende' zweifellos eine große Chance, da sie die Entwicklungs- und Deutungsmöglichkeiten beider Disziplinen auf eine neue Grundlage stellt; die methodischen Zwänge und die Schwerfälligkeit, die sich aus der Einführung informatischer Elemente ergeben, sind jedoch nicht zu unterschätzen und müssen in der Planung der universitären Ausbildung wie einzelner Projekte mit der nötigen Vorsicht berücksichtigt werden.

Für die universitäre Lehre ergibt sich daher langfristig die Notwendigkeit, ein Zusammenspiel von programmiertechnischen Elementen und traditionellen Kenntnissen in Paläographie, älteren Sprachstufen sowie der Editionspraxis zu gestalten.

[18] Nur ein Beispiel: Das XML-Schema zur Beschreibung von XML-Dokumenten, das entscheidende Vorteile gegenüber der älteren DTD aufweist (größere Präzision, Begrenzung der Wahlmöglichkeiten für bestimmte Felder [und damit Fehlerreduktion], Kodierung in XML) hat sich seit seiner Entwicklung (s.o.) in kaum fünf Jahren als neuer Standard weltweit und in allen Anwendungsbereichen durchgesetzt.